Variance Analysis of Complete Designs

Some Practical Aspects

By Paul Seeger

ALMQVIST & WIKSELL · STOCKHOLM

Uppsala 1966

Variance Analysis of Complete Designs

Some Practical Aspects

BY

Paul Seeger
Fil. lic., Smål.

INAUGURAL DISSERTATION

BY DUE PERMISSION OF THE FACULTY OF SOCIAL SCIENCES
OF THE UNIVERSITY OF UPPSALA
TO BE PUBLICLY DISCUSSED IN LECTURE ROOM IV
ON OCTOBER 8TH, 1966, AT 10 A.M.,
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

UPPSALA 1966
ALMQVIST & WIKSELLS BOKTRYCKERI AB

Variance Analysis of Complete Designs

Variance Analysis of Complete Designs

Some Practical Aspects

BY

Paul Seeger

UPPSALA 1966

Printed in Sweden by

Almqvist & Wiksells

Boktryckeri Aktiebolag

Uppsala 1966

PREFACE

Much has been said about the analysis of variance; some may think that everything has been said. See, for example, the extensive bibliography referred to only in this monograph. Analysis of variance is treated in most text books on statistical theory and applications, and there are special text books solely devoted to this subject. However, it is not always easy to find the most suitable way of dealing with a specific problem in a specific situation. The text book may be difficult to read because it is written in an abstract manner, aiming at elegant mathematics, or it may be of the cook-book-type, giving little insight into the method. Some problems can be solved in many different ways, but it is not self-evident which is the best. Sometimes assumptions and hypotheses are inadequately examined or insufficiently discussed.

This book tries to bridge the difficult gap between theory and practice, at least for a consulting statistician, but even for interested investigators, who are not professional statisticians. It will present my collected experience from practical applications with analysis of variance, and explain how and why I use the method. This means, of course, that the book will be limited to what I have met with in practice, and to what has specially interested me.

The applicability of the analysis of variance will be discussed in connection with a large number of real examples, and some theoretical results as well as data from simulation experiments will be given. I have endeavoured to find out what happens to the analysis using the most realistic assumptions: when inconvenient interactions are introduced and when the number of observations is a random variable etc. Great care has been taken to specify populations, models and hypotheses, especially for the methods used instead of the analysis of variance when the measurements are at a weak level; that is, when the nominal and ordinal scales are used. The classical assumptions for the analysis of variance are not satisfied when using these "weak scales".

The discussion is held at a rather elementary level, especially in part I, but to some extent even in parts III and IV. No special training in analysis of variance above that given in a general course in statistical analysis should be necessary to understand these parts. Some paragraphs in part II

are more difficult to follow. Brownlee's "Statistical Theory and Methodology in Science and Engineering", or similar literature corresponding to it should, however, facilitate the reading.

Throughout my studies of the analysis of variance and the writing of this thesis, I have had the opportunity of having many valuable discussions with Professor Herman Wold. He has always stressed the importance of a clear presentation of theoretical and applied work, and has taken a special interest in the comparison between regression analysis and the analysis of variance; a comparison that he has further elucidated with the help of the predictor concept. For all this I want to express my deepest gratitude.

My gratitude is extended to Docent Gunnar Eklund for stimulating and informative discussions especially concerning the problems in parts III and IV, to Docent Ejnar Lyttkens, who is always willing to discuss any statistical problem, and to the members of the seminar at the Institute of Statistics in Uppsala, to whom I have presented all parts of the thesis.

I also wish to thank Fil. lic. A. Gabrielsson, for kindly providing me with practical problems and data, Mr. J. Sandzelius, for helping me with the linguistic aspects, and my wife for reading through and criticizing the several versions of the manuscript and for drawing the figures.

Finally I want to thank Miss M. Laadi, Miss E. Pettersson, Mrs. V. Rannaleet, Miss A. Rask and Mrs. V. Svärd for typing the manuscript.

Uppsala, May 1966

Paul Seeger

CONTENTS

Part I. Introduction

1. The analysis of variance and its relationship to regression analysis and to sampling 3
   1.1. What analysis of variance means 3
   1.2. Relationship to regression analysis 6
      1.2.1. Introduction
      1.2.2. Formal similarities
      1.2.3. Relationship when the methods are applied to the same problem
      1.2.4. Conclusions
   1.3. Relationship to sampling 13
2. Some historical notes 17
3. Scope of this monograph 21
   3.1. Different types of design 21
   3.2. A summary of the following parts 22
      3.2.1. Part II
      3.2.2. Part III
      3.2.3. Part IV

References 25

Part II. General problems of analysis

1. The choice of model 29
   1.1. Introduction 29
   1.2. Population models and statistical models 30
   1.3. Applications 35
   1.4. Conclusions 41
2. Estimation of variance components 43
   2.1. Introduction 43
   2.2. Balanced samples 43
      2.2.1. A small introductory example
      2.2.2. Expected mean squares
      2.2.3. The sampling case
      2.2.4. The experimental case

2.2.5. Discussion
2.2.6. Applications
2.3. Unbalanced samples 61
2.3.1. Introduction
2.3.2. Some notes about the literature on unbalanced situations
2.3.3. Modified analysis with a randomly unbalanced sample
2.3.4. Conclusions
2.3.5. Applications
2.4. Notes about the reliability of the estimates 81
3. Testing hypotheses 82
3.1. Introduction 82
3.2. Problems in the classical normal theory 83
3.2.1. Inexact tests with mixed models
3.2.2. No suitable mean square for error
3.3. General models with few assumptions 88
3.3.1. Illustration of a permutation test
3.3.2. Normal approximations
3.3.3. An application
3.3.4. Numerical values for the correction factor
3.3.5. t-tests for pairs
3.4. Conclusions 98
4. Individual comparisons 100
4.1. Introduction 100
4.2. Choice of contrasts 101
4.3. Derivation of estimates 103
References 107

PART III. THE PROBLEM OF MULTIPLE COMPARISONS

1. Statement of the problem 111
2. Mass-significance in connection with F-tests 114
2.1. The case of many factors 114
2.2. The case of many levels per factor (or factor combination) . . 115
3. Different operational definitions of type-I-error frequencies 117
4. Reviews and illustrations of existing methods 120
4.1. Methods with constant least significant difference 120
a) The S-method
b) The T-method
c) The multiple-t- and LSD-methods
d) Fisher's method

4.2. Methods with variable least significant difference 125
a) Newman-Keuls' test
b) Tukey's method from 1953
c) Duncan's new multiple range test
4.3. Method for comparisons with a standard 127
Dunnett's method
4.4. Some non-parametric methods 128
a) The sign test
b) t-tests with individual estimates of variances
c) Rank tests
4.5. Two older methods 132
a) Tukey's method from 1949
b) Duncan's new multiple comparisons test from 1947
4.6. A discussion of Eklund's method 134
5. Practical recommendations with further numerical illustrations . . 142
5.1. Designs of a moderate size 142
5.1.1. Contrasts involving more than two treatments
5.1.2. Pairwise contrasts
5.1.3. Comparisons with a standard
5.1.4. Standard assumptions not fulfilled
5.2. Very large designs. 148
5.3. Summary 156
References 158

Part IV. Scale problems

1. Introduction 163
2. Nominal scale 165
2.1. Concerning the scale. 165
2.2. Several related samples (Randomized block designs) 165
2.2.1. Introduction
2.2.2. An application
2.2.3. Cochran's Q-test and an extension to it
2.2.4. Another derivation of the test formula
2.2.5. The possibility of varying p-values
2.2.6. Different hypotheses and models
2.2.7. The possibility of using F-tests
2.2.8. A numerical example
2.2.9. Some simulation experiments
2.3. Summary 190

3. Ordinal scale 191
3.1. Introduction 191
3.2. Completely randomized designs with one factor 191
3.3. Randomized block designs with one factor 193
3.3.1. Two levels
3.3.2. More than two levels
3.4. More than one factor 222
References 225

Part I

INTRODUCTION

CHAPTER 1

THE ANALYSIS OF VARIANCE AND ITS RELATIONSHIP TO REGRESSION ANALYSIS AND TO SAMPLING

1.1. What analysis of variance means

The analysis of variance is often identified by the F-test. Hypotheses about one or two mean values are said to be tested by a λ-test or a t-test, while hypotheses about at least three mean values are tested by analysis of variance. Of course, this is not a completely correct description. Analysis of variance is better defined as a method of dividing (analysing) a total variance into at least two components, which are perhaps synthesized into other variances. Using a liberal definition of variance (so liberal, for example, that $(\bar{x}-\mu)^2/1$ is called a variance) analysis of variance can be said to consist of estimating variances, and possibly of testing hypotheses about them. This "definition" is the same as Scheffé's (24) verbal explanation of what is meant by analysis of variance.

A few terms to be used below might need an introduction. The term "treatment" originates from agricultural experiments in which different plots of land are treated in different ways, but is now used for any conditions to be compared. A group of similar treatments (a factor) is said to have a number of "levels". The term "level" also has an agricultural origin; a group of similar treatments could consist of different levels or amounts of fertilizer used. Treatments and levels are often used as synonyms.

In any application it is always possible to construct a model for the observed values, a model reflecting the different sources of variation which are (believed to be) operating. For example, if one measures the percentage of alcohol in a homogeneous liquid using two different methods (treatments); the jth observation of the method i may be written:

$$y_{ij}=\mu+\alpha_i+e_{ij}; \qquad \begin{matrix} i=1,\ 2 \\ j=1,\ 2\ \ldots\ r_i \end{matrix}$$

where μ is the true percentage, α_i the systematic error of the ith treatment, and e_{ij} the error of measurement. For mathematical reasons it is common to define the constants μ and α_i so that $\Sigma_i \alpha_i=0$. Then the α_i need no longer express systematic errors. However, μ is the mean value of the

treatment means, α_i reflects differences between methods, and e_{ij} random variations within the methods. The total variation between the observations y_{ij} can be analysed into two components, variation between and within methods.

The model used here will be called Model I of the analysis of variance. The characteristic feature of this model is that our interest lies only in the (two) methods in the experiment; this is expressed in the equation of the model by the fact that the treatment effects α_i are constants. If the treatments in the experiment are considered as a sample from some population, about which one wants to generalize, the model will be called Model II. Assume that we want to determine the percentage of Fe_2O_3 in a vessel with "smältferrit" (a ferro-magnetic material), which is suspected of being unhomogeneous as regards Fe_2O_3. Assume further that a samples[1] from randomly chosen places in the vessel are taken, and that r analyses are made on each sample. (After the sampling each sample was made homogeneous.) The equation of the model is then written as:

$$y_{ij}=\mu+a_i+e_{ij}; \quad \begin{matrix} i=1,2...a \\ j=1,2...r \end{matrix}$$

The treatment effects (reflecting differences between samples) are now random variables denoted by Latin letters. In Model II it is more meaningful to talk about two components of the variance, in this example the variance between samples within the vessel and the variance of observational errors.

When the experiment includes two or more different groups of treatments, for example, if for every sample of "smältferrit", b analysts make r determinations of the percentage of Fe_2O_3, the difference between the two models becomes more accentuated and in addition we are forced to introduce a so called mixed model. When studying possible differences between the b analysts (and between no other analysts) the equation of the model becomes:

$$y_{ijk}=\mu+a_i+\beta_j+e_{ijk}; \quad \begin{matrix} i=1, 2 \ldots a \\ j=1, 2 \ldots b \\ k=1, 2 \ldots r \end{matrix}$$

where a_i reflects true differences between samples, β_j true differences between analysts, and e_{ijk} the measurement error. This experiment is said to include two factors, namely:

[1] Note that "sample" is here used in another sense than the usual "sample from a population" sense.

1) samples with a levels,
2) analysts with b levels.

The factor "samples" is random while the factor "analysts" is fixed, thus we have a mixed model.

If there is reason to suspect an interaction between the two factors (the differences between the analysts could vary with the percentage of Fe_2O_3 in the samples) an interaction term is introduced in the equation, which can be written as follows:

$$y_{ijk}=\mu+a_i+\beta_j+(a\beta)_{ij}+e_{ijk};\quad \begin{matrix} i=1, 2 \ldots a \\ j=1, 2 \ldots b \\ k=1, 2 \ldots r \end{matrix}$$

The various assumptions made about the different effects in the model will be discussed in detail in Part II. This much can, however, already be said that even for two factors there are several variance components to estimate and several hypotheses to test. How this will be done depends on which model turns out to be the correct one.

Examples of more or less complicated variants of these three basic models are given below.

As mentioned earlier it is a common view that analysis of variance is a technique for coping with three or more means. Using the definition that it consists of analysing a total variance into at least two components, the area of application will thus be extended. Assume, for example, that we want to test a hypothesis $H_0 : \mu=\mu_h$ about the mean in a normal population, and that we investigate a random sample $y_1, y_2 \ldots y_r$ from this population. The "total variation" can then be split into two components in the following way:

$$\sum_{i=1}^{r}(y_i-\mu_h)^2=\sum_{i=1}^{r}(y_i-y_.)^2+r(y_.-\mu_h)^2.(^1)$$

The equation of the model can then be written as:

$$y_i=\mu_h+(\mu-\mu_h)+e_i,$$

and the traditional analysis of variance table will have the following appearance:

---

[1] When a subscript is replaced by a dot all the observations corresponding to different values of that subscript have been averaged.

Sources of variation	SS = sums of squares	D.f. = numbers of degrees of freedom	MS = mean squares	E(MS) = expected mean squares
Between	$r(y_{\cdot} - \mu_h)^2$	1	$r(y_{\cdot} - \mu_h)^2$	$\sigma_e^2 + r(\mu - \mu_h)^2$
Within	$\sum_{i=1}^{r} (y_i - y_{\cdot})^2$	$r-1$	$\frac{1}{r-1} \sum_{i=1}^{r} (y_i - y_{\cdot})^2$	σ_e^2
Total	$\sum_{i=1}^{r} (y_i - \mu_h)^2$	r		

In this table the variance σ_e^2 of the normal population has been introduced.

Generally the analysis of variance is said to consist of the following steps:

1) construction of the relevant model for the particular investigation;
2) partitioning the total sum of squares and the total number of degrees of freedom into components corresponding to the effects in the equation of the model;
3) using the mean squares, that is the quotients between the sums of squares, and numbers of degrees of freedom for the different components, for estimation and perhaps tests of hypotheses.

In order to be able to make the estimates, one has to compute the expected mean squares, the E(MS)'s. Much space will be devoted to the derivation of formulas for E(MS) in this paper. Tests of hypotheses about variances will as a rule require special assumptions about distributions. However, the F-test has proved itself workable enough in most practical cases.

1.2. Relationship to regression analysis

1.2.1. *Introduction*

In elementary courses in statistics the student is taught two important methods of analysis, regression analysis and the analysis of variance, which he finds quite different. The analysis of variance is given as an extension of the t-test, and regression is used as a means of fitting a straight line to bivariate data.

When the student starts to study more independently, he finds, when skimming the common text books in statistical theory and methodology, that the analysis of variance is given considerably more space than regression analysis. Upon closer study he reads, for example, in Ostle (21):

"Analysis of variance and regression analysis are essentially the same." The reason why the analysis of variance is treated so thoroughly is said to be because "there are many cases (based on specific conditions) that are more easily explained using the methods of this and succeeding chapters" (that is the chapters on analysis of variance).

Ostle is also referring to Chew (2), where one can read: "Analysis of variance and regression analysis are essentially identical." Two pages of theoretical derivations (of course completely correct in detail) are said to show this. Scheffé says in his book (24): "in the analysis of variance all factors are treated qualitatively, in regression analysis all factors are quantitative and treated quantitatively, whereas in the analysis of covariance some factors are present that are treated qualitatively and some that are treated quantitatively".

Soon the student of statistics finds that the opinion about the relative importance of the two methods of analysis differ considerably among statisticians. Some of them hold the view that the regression analysis is more general, and perhaps, that the analysis of variance is only an unimportant special case, while others hold that analysis of variance is an independent method of greater practical value. The only sense our student can make of this is that he has to build up an opinion of his own.

In the following few pages some similarities and differences as regards techniques as well as applications will be mentioned.

1.2.2. *Formal similarities*

In regression analysis, analysis of variance and covariance the problems treated can be summarized in the model:[1]

$$\mathbf{y} = \mathbf{x}'\boldsymbol{\beta} + \mathbf{e}$$

where:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix}$$

stands for the observed values of the dependent variable, and

[1] Other models, for example multiplicative ones, are sometimes used in regression analysis. They may, however, often be approximated by polynomials. Only so-called linear models will be treated below.

$$\mathbf{x} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ x_{p1} & x_{p2} & & x_{pn} \end{bmatrix}.$$

for the observed values of the independent variables $x_1, x_2 \dots x_p$. $\boldsymbol{\beta}$ is the vector to be estimated and $\mathbf{e}$ the residual.

If the independent variables are quantitative we have a regression problem; when they are indicator variables (possible values 0 and 1) there exists a situation suitable for the analysis of variance. For example, let us look at the model equation:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}; \quad \begin{matrix} i = 1, 2 \\ j = 1, 2 \end{matrix}$$

With the help of the indicator variables $x_0 - x_8$, we can write this as:

$$y_{ijk} = \mu x_0 + \alpha_1 x_1 + \alpha_2 x_2 + \beta_1 x_3 + \beta_2 x_4 + (\alpha\beta)_{11} x_5 + (\alpha\beta)_{12} x_6 + (\alpha\beta)_{21} x_7 + (\alpha\beta)_{22} x_8 + e_{ijk}$$

For y_{11k} we then have: $x_0 = x_1 = x_3 = x_5 = 1$, and

$$x_2 = x_4 = x_6 = x_7 = x_8 = 0.$$

This formal relationship between "additive" regression models and analysis of variance models has been commented upon by several authors; see for example: Brownlee (1), Chew (2) and Scheffé (24).

If the estimation vector $\boldsymbol{\beta}$ is fixed, as in Model I of the analysis of variance, there is a satisfactory theory behind the use of least squares methods for estimating $\boldsymbol{\beta}$. In this case we can say that the methods are essentially the same, and that the analysis of variance is more easily computed because the x-values are restricted to 0 and 1. However, it should be observed that the equality is purely formal; different vectors $\boldsymbol{\beta}$ are estimated even in the same application.

When the vector $\boldsymbol{\beta}$ is random (Model II), or contains both fixed and random components (mixed model), we are outside the domain of least squares and the formal equality vanishes. However, the sums of squares derived by least squares methods in Model I are often used in Model II

and in the mixed model, even if we do not know whether or not they give optimal estimates, as they can be proved to give in the fixed case.

The model equation written as in regression analysis would be:

$$y = \beta_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p + e$$

where $x_1 \ldots x_p$ are indicator variables. The problem is to estimate the mean matrix and the covariance matrix for $\mathbf{b}' = (b_1 \ldots b_p)$.

It would be possible to speak about an extended regression problem if we allowed the independent variables to be quantitative.

As an example consider the model equation:

$$y = a + bx + e,$$

where a, b, e are independent random variables with means $\alpha\,;\,\beta\,;\,0$ and variances $\sigma_a^2;\ \sigma_b^2;\ \sigma_e^2$. If the x-values are predetermined we have:

$$E(y \mid x) = \alpha + \beta x\,;\ D^2(y \mid x) = \sigma_a^2 + \sigma_b^2 x^2 + \sigma_e^2.(^1)$$

In this case the estimation problem can obviously not be solved without making further restricting assumptions. An application is hinted at on page 11.

1.2.3. *Relationship when the methods are applied to the same problem*

As has been pointed out by Scheffé (24), regression analysis and the analysis of variance have different domains of applications. When the independent variables are qualitative, for example: methods of analysis, varieties of wheat etc., we apply analysis of variance; a regression curve is meaningless when there is no natural order between the factor levels.

When quantitative variables are used, we can as a rule extract more information from the data if we use regression analysis, but sometimes we treat the variables qualitatively and use the analysis of variance.

Let us discuss some similarities and differences when there is only one quantitative independent variable x.

Model I. Four values, x_1–x_4, have been chosen for the independent variable x, and in the analysis of variance we are interested in just these four values, and we want to test the hypothesis:

[1] In conformity with current usage we specify regression models in terms of conditional expectations (= predictors); see e.g. Kendall: The advanced theory of statistics, Plackett: Regression analysis, and Wold: Forecasting by the chain principle, paper in Econometric model building: The causal chain approach, North-Holland Publ. Co., Amsterdam (1964).

$$H_0: E(y_{ij}|x_i) = \mu; \quad i = 1,2,3,4$$

against the alternative that not all four means are equal.

We could also use regression methods to estimate $E(y_{ij}|x_i)$ as a function of x, for example, estimate α and β in the linear model:

$$E(y_{ij}|x_i) = \alpha + \beta x_i.$$

The hypothesis:

$$H_0: \beta = 0$$

corresponds to the analysis of variance hypothesis above. However, the alternative here is that the mean values lie on a straight line with $\beta \neq 0$. For the first test we have 4-1, and $\Sigma_i r_i - 4$ degrees of freedom, for the second 1, and $\Sigma_i r_i - 2$, which reflects the fact that we have a more restricted model in regression analysis. Note that the estimation vector in the first case is $(\mu_1; \mu_2; \mu_3; \mu_4)$, and in the second $(\alpha; \beta)$, interrelated by: $\mu_i = \alpha + \beta x_i$;.

Model II. In experiments with a quantitative factor Model II is sometimes used even when the levels are predetermined; they are considered as representative of neighbouring values. Here we will discuss a more clear-cut Model-II-situation.

We are interested in the x-values in a certain interval $x' \leqslant x \leqslant x''$. Data for the analysis of variance is then collected in the following way: We choose at random a x-values in the above-mentioned interval, and then observe a certain number of y-values for each of these x-values; the same a x-values are used all through the experiment. The model equation can be written as follows:

$$y_{ij} = \mu + a_i + e_{ij}; \qquad \begin{matrix} i = 1, 2 \ \ldots \ a \\ j = 1, 2 \ \ldots \ r_i \end{matrix}$$

The main problem is to estimate σ_a^2, the variance between $E(y_{ij}|x_i)$, for all possible (evenly distributed) x_i's in the interval $x' \leqslant x \leqslant x''$.

In a regression problem the data could be collected in the same way as above (that is, by sampling y-values for the a sampled x-values), and then the comparison between regression and the analysis of variance would follow the same lines as in Model I. Regression methods can also be applied to a bivariate random sample $(y_i; x_i)$, possibly with different x-values for every pair. In this way regression analysis could be said to be more general. If we should apply analysis of variance to a bivariate random sample we

would run the risk of observing only one y-value for every x (with r_i a random variable), in which case σ_e^2, and thus σ_a^2, could not be estimated. The corresponding estimations are possible in a regression problem, because we have introduced a more restricted model, for example, the model $E(y|x) = \alpha + \beta x$, which gives us the opportunity to estimate σ_e^2 from residuals around the line, instead of from within groups.

The hypothesis:

$$H_0: \sigma_a^2 = 0$$

(which is seldom tested) is then equivalent to:

$$H_0: \beta = 0$$

for a linear regression model, or to:

$$H_0: \beta = \gamma = \delta = 0$$

for the model: $E(y|x) = \alpha + \beta x + \gamma x^2 + \delta x^3$. In this application the analysis of variance model:

$$y_{ij} = \mu + a_i + e_{ij}$$

or, written as in paragraph 1.2.2:

$$y_{ij} = \mu x_0 + a_1 x_1 + a_2 x_2 + \ldots + a_a x_a + e_{ij}$$

has a random estimation vector, while the regression model:

$$y_{ij} = \alpha + \beta x + \ldots + e_{ij}$$

has a fixed estimation vector. It will be noted that the analysis of variance model, too, can be written in predictor form[1],

$$E(y_{ij}|x_1 \ldots x_a) = \mu x_0 + a_1 x_1 + \ldots + a_a x_a$$

As mentioned before, it would be possible, at least theoretically, to extend the regression problem to a random estimation vector. For example, suppose that we want to compare some different types of rain-collecting gauges. We know that there are not only differences between types of gauges, but also between gauges of the same type. Besides, these differences vary with the amount of precipitation (and perhaps with other factors neglected here). For the sake of simplicity we may assume that for every gauge the collected amount of precipitation is a linear function of the true

[1] Reference is made to prof. Wold's treatment of analysis of variance in a text book in preparation.

amount of precipitation. If a number of gauges of a certain type are chosen at random from "the population of gauges of that type", we can write an observed amount of precipitation for gauge i, and rain j like this:

$$y_{ij} = a_i + b_i x_j + e_{ij}$$

where x_j is the true precipitation. In the population there are a number of pairs of parameters: $(\alpha_i; \beta_i)$; $i = 1, 2 \ldots N$, with means $(\alpha_-; \beta_-)$, and variances $(\sigma_\alpha^2; \sigma_\beta^2)$; so we can say that a_i and b_i are (dependent) random variables, having the above-mentioned means and variances. By comparing two types of gauges A and B it could be of interest to estimate:

$$\int_x \{(\alpha_-^A + \beta_-^A x) - (\alpha_-^B + \beta_-^B x)\}\, dx,$$

where the integration is for "normal rains". One of the many difficulties is that x cannot be measured without error.

1.2.4. *Conclusions*

On the basis of this discussion we can now conclude that there is a formal equivalence between regression analysis and the analysis of variance in Model I. Both methods apply least-squares-methods to models of the type:

$$\mathbf{y} = \mathbf{x}'\boldsymbol{\beta} + \mathbf{e},$$

in which the x-values are more general in the regression situation. However, different parameters, $\boldsymbol{\beta}$, are estimated with the two methods.

Turning to Model II in the analysis of variance we notice that there is no longer any formal equivalence; as random estimation vectors are outside the ordinary regression analysis. In one and the same practical problem however, Modell II corresponds to a fixed parameter vector, $\boldsymbol{\beta}$, in regression.

In constructing the regression model we often utilize some a priori information and obtain thereby more information a posteriori. The analysis of variance model, on the other hand, is completely determined by the experiment (or sampling investigation). However, in planning the investigation we have as a rule utilized some a priori information of a different nature than that used in determining the form of the regression model. Thus we can say that a priori information is used differently in connection with regression analysis and the analysis of variance.

A more important difference between the two methods is that they are applied in different situations: analysis of variance when the x-variables are qualitative, and regression analysis when they are quantitative. Some-

times quantitative variables are treated qualitatively and the variance analysis is then applied. As is shown in most text books there is a correspondence between the parameters in the analysis of variance and the regression models; we can therefore start with an analysis of variance and perhaps follow up with a regression analysis. Such methods are used when there are several quantitative variables and we want to find optimal profit (response-surface-technique).

In practice there are often both qualitative and quantitative variables in an experiment; we are then interested in both the analysis of variance and the regression analysis. Hence it is hardly correct to claim that one method is more general than the other.

1.3. Relationship to sampling

The student of statistics may also get a feeling that there is some relation between sampling and the analysis of variance. He finds that the analysis of variance is discussed in text books on sampling, and he hears older colleagues say that if you are familiar with the analysis of variance then you know a lot about sampling.

Of course one can say that they can not be compared on the same level. Sampling is a sub-discipline of statistics, just like experimental design, demography, econometrics etc., while the analysis of variance is a method of analysis used in some of these sub-disciplines.

However, sampling consists of methods of making intelligent data choices, of constructing good estimators for various parameters in the population, and of deriving and estimating variances for the estimators. For this latter estimation the analysis of variance is used.

A large part of the literature on the analysis of variance emphasizes F-tests and Model I. In a new text book by Guenther (17), one can read: "In most applied work where a one-way classification is used, Model I is the appropriate choice." Perhaps authors concentrating on tests have found their practical problems in the field of experimental design.

The two main fields of application for the analysis of variance are experimental design and sampling. In experiments it may often be true that the factor of interest is fixed; we compare, for example, different varieties of wheat, methods of analysis, types of washing-machines etc. It is then that tests come to the foregrund. Random factors may be included, but as a rule we are only interested in eliminating their variances. It is also characteristic of experiments to have well defined experimental units, such as individuals, animals or field plots, to be treated differently, and to distribute

these units among the treatments either through complete randomization or randomization restricted in some way.

In sampling situations, however, the factors are random as a rule, and the interest lies in estimating variances. An example has already been mentioned in paragraph 1.1, namely sampling from a vessel of "smältferrit" in order to determine the average percentage of Fe_2O_3 occurring. Another example is provided by an investigation made in order to estimate the average yield per ha.[1] of wheat in a certain district. We could then sample farms, fields within farms, and identically treated plots. The variance analysis problem is then to separate the variances between farms, between fields within farms and between plots within fields. As is often the case in sampling investigations we have a hierarchical (nested) composition, plots are nested within fields, which are nested within farms. In a randomized block experiment, for example, blocks and treatments are said to be crossed, as treatment no. 1 is the same in every block. This is contrary to the case of the nested investigation, where field no. 1 on one farm is different from field no. 1 on another farm.

Sometimes the numbers of observations are random variables. For instance, if, in the above example, the farms vary in size, it would be natural to sample a constant fraction of the fields per farm. Thus the number of fields to be investigated depends upon which farms have been chosen. Another modification of the analysis of variance is needed if we are sampling with unequal probabilities.

The border line between experimental designs and sample surveys is quite diffuse; the factor levels in an experiment may be sampled from a population of levels, and many sampling investigations could just as well be called planned experiments.

Further aspects on the distinction between the sampling case and the experimental case can be obtained by studying the fourfold table of statistical methods introduced by Wold (31,32). The purpose of a statistical investigation can be descriptive or explanatory, and the data experimental or non-experimental. Using this twofold dualism and introducing intermediate cases, the fourfold table is reproduced in Table I: 1.

The experimental case belongs to the SW-region. An example can be provided by comparing varieties of wheat on different plots of land. When only one variety is grown on equivalent plots the purpose must be purely descriptive. However, the above does not constitute an application that is to be analysed by the analysis of variance, unless we have a hypothesis

[1] 1 ha., an abbreviation of "hektar", is about equal to 2.5 acres.

Table I: 1.

Purpose	Experimental data	Intermediate	Non-experimental data
Description	A. Replications under constant conditions B. Independent replications Sampling: $S(\rightarrow P)\rightarrow U$ **NW**	A or B distorted	A. Constant conditions distorted Stratification B. Independence distorted. Sampling: $S\rightarrow P(\rightarrow U)$ **NE**
Intermediate	Explanatory model: Replications under constant conditions		Symmetric treatment of variables: $r(x, y)$, $\alpha(x, y)$ Testing for association: χ^2
Explanation	Explanatory model: **SW** $x\rightarrow y$ x steered Test-control group experiments Stimulus-response experiments In each group: Independent replications under constant conditions	$x\rightarrow y$ Some x steered, some x not Operations analysis	**SE** Explanatory model: $x\rightarrow y$ x not steered Other influences: Specification errors

of the average yield, in which case a formal partitioning of a total variance into two components is possible (see page 5). We shall then obtain an intermediate case between description and explanation.

The sampling case, however, cannot be placed in only one region. Let us look at the examples mentioned above and see if we can classify them. If we want to estimate the average percentage of Fe_2O_3 in a vessel with "smältferrit", the purpose is descriptive. The same applies if it is the variation between methods, between analysts or within vessels which interests us. However, we may also want to explain observed differences between methods or analysts. The data are intermediate between experimental and non-experimental as it is possible to make a few analysis on the same homogeneous sample of "smältferrit" but not to replicate the sample. More clear-cut NW- or SW-data are obtained from analyses on standard solutions that can be replicated exactly. If we want to determine the average yield, or the variations in yield between farms and fields, in a certain district, we have an application in the NE-region. If the purpose

is to explain differences in yield by soil conditions, farmer's ability etc. we are in the SE-region, and if we want to compare the precision of two different methods of determining the average yield, we are in the SW-region.

Thus the sampling case includes many more different situations than the experimental case and is also more common in practice.

CHAPTER 2

SOME HISTORICAL NOTES[1]

The analysis of variance originates from R. A. Fisher in about the year 1920.[2] In 1919 Fisher came to the Rothamsted experimental station and was faced with the problem of estimation of errors. The state of agricultural research at that time is described by Yates (33): "What was almost completely lacking was any coherent theory on the estimation of errors from the results themselves, except in the simple case of a comparison of two treatments only." Yates goes on to talk about the analysis of variance: "Certainly it immediately clarified the problem of estimation of experimental error."

Fisher (11) published a paper as early as 1921, in which he examined yield records from 13 plots, uniformly treated for 67 years. Using orthogonal polynomials he fits curves to the yields for every plot, and divides the total variation over the years in "(1) annual variation (2) steady diminution due to deterioration of the soil (3) slow changes other than steady diminution". The additive property of variances "enables us to *analyse* the total, and to assign, with more or less accuracy, the several portions to their appropriate causes, or groups of causes". Then he shows: "the analysis of the total variance for each plot". In 1923 "Student" (26) quotes a letter from Fisher in which he presents the analysis suitable for randomized blocks. This was an example of the simplicity and elegance of the analysis of variance as a tool for estimating error variances. "Student" had tried a complicated method using pooled differences between every pair of treatments within blocks, a method which turned out to be the same as Fisher's. In the same year Fisher and Mackenzie (16) published a comparison between different varieties of potatoes in a split-plot design. The analysis

[1] Most of following notes and references to early original papers have been collected from papers by Eisenhart (7), Scheffé (23), Plackett (22), Mahalanobis (20), Yates (31) and Finney (8).

[2] The Model-I-problem, being an ordinary least-squares-problem, could be said to be as old as the method of least squares, invented at the beginning of the 19th century. Scheffé (23) also gives references to astronomers making explicit use of Model II in the 1860's. However, I think it is correct to call Fisher the founder of the analysis of variance that is used today.

used in this latter paper was more like the present analysis of variance than was the partitioning of the total variance in (11). Therefore (16) is referred to by Snedecor (25) as the paper in which Fisher introduced the analysis of variance.

The theoretical tool for making significance tests in the analysis of variance, the z-test, was developed by Fisher in his studies on the distribution of correlation coefficients. His paper of 1918, (10), is often mentioned as the birth of the analysis of variance. In this paper Fisher introduces the terms "variance" and "analysis of variance" in his attempt to divide the total variance between (for example) the stature of brothers into portions ascribable to different causes. "For stature the coefficient of correlation between brothers is about 0.54 which we may interpret by saying that 54 per cent of their variance is accounted for by ancestry alone, and that 46 per cent must have some other explanation." "... we will attempt to determine how much more of the variance, in different measurable features, beyond that which is indicated by the fraternal correlation, is due to innate and heritable factors."

The z-transformation of correlation coefficients was discussed in papers of the years 1915 (9), 1921 (12), while the general z-distribution was developed in 1924 (13). With the help of z, the distribution of the intraclass correlation coefficient for pairs was transformed to an approximately normal distribution almost independently of the number of pairs. Even for a more general comparison between two variances or mean squares the distribution became approximately normal, if the numbers of degrees of freedom, n_1 and n_2, were very large. However, for small and moderate numbers the approximation may be very bad, and therefore Fisher derived the exact distribution of z and tabulated it in his "Statistical methods for research workers" 1925 (14).

It was soon realized that it would be more convenient in the analysis of variance to compute the quotient of mean squares, that is, e^{2z} instead of z. Mahalanobis (19) seems to have been the first to calculate it, and in 1934 Snedecor (25) named it "F" in honor of Fisher and tabulated it after Fisher's above-mentioned tables. It may be interesting to quote from Snedecor's book: [by] "calculating the ratio of the larger mean square to the smaller. [and] Designating this ratio by F,

$$\text{[we get:]} \quad F = \frac{\text{larger mean square}}{\text{smaller mean square}},,$$

The habit of taking the larger variance in the numerator could become relevant when treating intraclass correlation, where a larger mean square

"within" than "between" suggests negative intraclass correlation, but hardly in ordinary analysis of variance.

From the beginning Fisher used both fixed-effects models and random-effects models (the latter, at least, when he treated intraclass correlations), and he mentions what has later been named a "mixed model" in "The design of experiments", 1935 (15).

The useful "analysis of variance table", including sums of squares, degrees of freedom and mean squares for the various sources of variation, was first published (as far as the author knows) in Fisher's paper together with Mackenzie in 1923, (16), and in "Student's" quotation of Fisher's letter of 1923, (26). Tippett in 1931, (28) appears to have been the first to include a column of expected mean squares for variance components and to estimate these components.

Eisenhart in 1947, (7) was the first to use the names "Model I" and "Model II" for fixed-effects and random-effects models, and he also mentions the possibility of a mixed model. Before that, more formal statements of the models than Fisher used had appeared in papers by Daniels in 1939, (6) and Crump in 1946, (4) as regards the random-effects models, and by Jackson in 1939, (18) as regards the mixed model. However, a more complete treatment of the mixed model did not appear until 1956, (23).

The inclusion of interaction effects goes back to Fisher in the very beginning of the analysis-of-variance-era. According to Scheffé (23), Crump in 1946, (4) was the first to clearly put into writing a random interaction effect. For a long time, factor effects and interaction effects were considered as independently and randomly sampled from different populations. Nowadays the assumed normality is inconsistent with the factor effects, a_i and b_j, being considered dependent but at the same time uncorrelated with the interaction effect $(ab)_{ij}$.

The development of various models was very intensive in the 1940's and 1950's. This development can be followed in expository articles by Crump in 1951, (5) and Plackett in 1960, (22). Crump considers the model equation (using his notation) as having the following form:

$$y_{hij} = \mu + \alpha_h + \beta_i + (\alpha\beta)_{hi} + \varepsilon_{hij}; \quad \begin{matrix} h = 1, 2 \ldots a \\ i = 1, 2 \ldots b \\ j = 1, 2 \ldots n \end{matrix}$$

in which "μ is a parameter which remains fixed in repeated samples, while the elements $\{\alpha_h\}$, $\{\beta_i\}$, $\{(\alpha\beta)_{hi}\}$ and $\{\varepsilon_{hij}\}$ are random samples of size a, b, ab, and abn respectively from populations π_α, π_β, $\pi_{\alpha\beta}$ and π_ε, each with mean zero, respectively". Six models with different assumptions

regarding the sizes and probability distributions were considered. Two of them are Eisenhart's Model I and Model II, the only difference is the interaction effect, and for Model I that the effects are obtained by exhaustive sampling from the populations π_α, π_β and $\pi_{\alpha\beta}$, instead of being fixed. The four other models were proposed by Tukey (29). One of them was something in between Model I and Model II; the effects were considered as sampled from finite populations.

According to Plackett (22), the main criticism against these models was that they did not take account of the practical execution of the experiment (or sampling); they only specified how the effects had been sampled. The emphasis lay on making the classical tests of significance valid. A serious drawback was the independent interaction. In practice, the interaction term, $(ab)_{ij}$, cannot be considered independent of the ith row and the jth column, that is of their effects a_i and b_j. Among those who took part in the construction of models, tied to the experimental or sampling situation, were Cornfield-Tukey (3) and Wilk-Kempthorne (30). Their various models will be discussed in part II. We can however, say that Cornfield-Tukey considered what is here called the sampling case, and Wilk-Kempthorne the experimental case. Plackett (22) characterized the difference between these approaches by saying that in Cornfield-Tukey's case, every factor combination had its own population of plots, while in Wilk-Kempthorne's all combinations, or groups of them, had a common plot population.

CHAPTER 3

SCOPE OF THIS MONOGRAPH

3.1. Different types of design

Design is used as a general term for the structure of observations. We distinguish between complete and incomplete, balanced and unbalanced designs. The meaning of these terms is illustrated by the following example already introduced in paragraph 1.1: b analysts are analysing a homogeneous samples of "smältferrit". As each analyst analyses the same a samples, we say that the factors $\mathcal{A}$ ("samples") and $\mathcal{B}$ ("analysts") are crossed, which can be visualized by looking at table I: 2.

Table I: 2. *The cell-frequencies in a two-way design with crossed factors.*

	$\mathcal{B}$					
$\mathcal{A}$	1	2	...	j	...	b
1	r_{11}	r_{12}		r_{1j}		r_{1b}
2	r_{21}	r_{22}				
⋮						
i	r_{i1}			r_{ij}		
⋮						
a	r_{a1}					r_{ab}

If all the cell-frequencies are positive, that is $r_{ij}>0$ for all i's and j's, the design is said to be complete, otherwise it is incomplete. If the cell-frequencies are equal, that is $r_{ij}=r>0$ for all i's and j's, the design is balanced, otherwise unbalanced.

The observations can be divided into quantitative and qualitative. (Examples of qualitative observations are counts and ranks.) Thus the field of application for the analysis of variance can be divided as in table I: 3.

This monograph is limited to balanced and unbalanced complete designs, which include three particular designs: 1) the sampling design (see paragraph 1.3), 2) the completely randomized (experimental) design and 3) the (experimental) randomized block design. Thus, the incomplete designs: Latin squares and cubes, Graeco-Latin squares, Youden squares, incomplete randomized blocks, lattices etc. are not treated.

Considering qualitative observations the discussion becomes most thorough for balanced randomized blocks.

An example with two crossed factors is given above. A factor is said to be nested within another if its levels differ from level to level of the other factor. For example the factor "fields" is nested within the factor "farms", because there are new fields for every farm. Thus we can distinguish between designs with crossed, nested and both crossed and nested factors. All these types will be considered.

3.2. A summary of the following parts

3.2.1. *Part II*

In chapter 1 the concepts[1] of population model and statistical model are explained and illustrated. In chapter 2 the estimation of variance components is discussed. Rules for writing down the expected mean squares in the general balanced case are given for the sampling case[2] and for completely randomized and randomized block designs in 2.2. In 2.3 a modification[3] of the unweighted analysis[4] is proposed. Expected mean squares are computed for random cell-frequencies, and it is suggested that, for crossed factors, the estimators for the variance components will be the same as in the balanced case (except of course for the fact that the mean squares are defined differently). An example of estimation when nested factors occur is also given.

Several problems in testing statistical hypotheses are discussed and extensively illustrated in chapter 3. It is, however, recommendable to use the simplest form of F-test in most cases where careful thought has not been able to reveal any extreme peculiarities in the population. Chapter 4 is concerned with the estimation of individual comparisons in the general case when variance inequality is allowed.

Ten examples drawn from the author's own experience of consulting work are used to illustrate the discussions throughout the chapter.

3.2.2. *Part III*

The first three chapters can be considered as giving an introduction to the problem of mass-significance. Four situations including many analyses

[1] Introduced by Wilk-Kempthorne (30).

[2] The rules in the sampling case are adapted from Cornfield-Tukey (3).

[3] Used by Sukhatme (27) for two-stage sampling.

[4] The usual least-squares-analysis uses cell means weighted with the number of observations in the sums of squares. However, it is easier and sometimes more relevant to use unweighted means, that is the unweighted analysis.

TABLE I: 3. *The field of application for the analysis of variance divided into different parts according to type of design and observation.*

Type of design	Type of observations: Quantitative	Type of observations: Qualitative
Complete — Balanced	Sampling investigations; Completely randomized experiments; Randomized block experiments — n treatment factors	Sampling investigations; Completely randomized experiments; Randomized block experiments — n treatment factors
Unbalanced	Ditto	Ditto
Incomplete	Latin squares and cubes, Graeco-Latin squares, Youden squares, Incomplete randomized blocks Fractional factorials etc.	For example tasting experiments with only two of the treatments per block (individual)

of significance are mentioned, one of which will be treated in the following. For this situation a list of 18 methods of solving the problem is drawn up and divided into groups according to various characteristics. Four different definitions of error rates are also presented.

In chapter 4 an inventory is made of the possibilities occurring in different cases (the 18 methods). Among the methods, the one by Eklund is discussed at length. As Eklund's method has an unusual definition of error rate, the probability of at least one significant result among N tests with true null hypotheses is derived.

Practical recommendations as to which methods are best are given in chapter 5. In very large designs most of the methods reviewed in the 4th chapter are shown to be too insensitive for use. A completely satisfying method for the case of a very large design is lacking.

Several examples from investigations comparing washing-machines, watches and psycho-physical reactions are used.

3.3.3. *Part IV*

The nominal scale (observations are counts) is treated in chapter 2, and the ordinal scale (observations are ranks) in chapter 3. Only the case of randomized blocks with one factor is examined when the nominal scale is used. All the while a particular psycho-physical application is kept in the foreground, but the treatment is nevertheless general. Cochran's Q-test is reviewed and extended to the case where the cell-frequencies are larger

than one. Two different developments of some relevance to the method of making simulation experiments are shown. Different null hypotheses and models are discussed, and the Q- and F-tests (with and without the arc-sine transformation) are compared in situations with real data from the psycho-physical application and in simulation experiments.

In chapter 3 some tests for the completely randomized design are discussed, but even here the discussion is concentrated on randomized blocks with one factor. When the factor has only two levels the relevant tests are the sign test and Wilcoxon's test. They are considered in detail, applying the most general assumptions. An endeavour is made through exemplification to clarify which null hypothesis the tests are really examining, and what the power may be in different situations. In the case of more than two factor levels the rank test of Friedman's is used. The same extension is made to this test as to that of Cochran's Q-test. The trouble caused by interaction between treatments and blocks is thoroughly investigated, and the possibility of making F-tests in order to avoid these troubles is discussed. A model like those used in part II is constructed, with functions of ranks as "effects".

In this section a distinction is made between quantitative and qualitative hypotheses and interactions. The examples used are from psycho-physics and meteorology.

REFERENCES

1. Brownlee, K. A.: Statistical theory and methodology in science and engineering. Wiley, New York (1960).
2. Chew, V. (ed.): Experimental designs in industry. Wiley, New York (1958).
3. Cornfield, J. and J. W. Tukey: Average values of mean squares in factorials. Ann. Math. Stat. 27 (1956), 907–949.
4. Crump, S. L.: The estimation of variance components in analysis of variance. Biometrics 2 (1946), 7–11.
5. — The present status of variance component analysis. Biometrics 7 (1951), 1–16.
6. Daniels, H. E.: The estimation of components of variance. J. R. Stat. Soc. Suppl. 6 (1939), 186–197.
7. Eisenhart, C.: The assumptions underlying the analysis of variance. Biometrics 3 (1947), 1–21.
8. Finney, D. J.: Sir Ronald Fisher's contributions to biometric statistics. Biometrics 20 (1964), 322–329.
9. Fisher, R. A.: Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika 10 (1915), 507–521.
10. — The correlation between relatives on the supposition of Mendelian inheritance. Trans. Roy. Soc. Edinburgh 52 (1918), 399–433.
11. — Studies in crop variation. I. An examination of the yield of dressed grain from Broadbalk. J. Agr. Sci. 11 (1921), 107–135.
12. — On the "probable error" of a coefficient of correlation deduced from a small sample. Metron 1 (1921), pt. 4, 1–32.
13. — On a distribution yielding the error functions of several wellknown statistics. Proc. Int. Math. Congr. Toronto. (1924), 805–813.
14. — Statistical methods for research workers. Oliver and Boyd, Edinburgh (1925).
15. — The design of experiments. Oliver and Boyd, Edinburgh (1935).
16. Fisher, R. A. and W. A. Mackenzie: Studies in crop variation, II. The manurial response of different potato varieties. J. Agr. Sci. 13 (1923), 311–320.
17. Guenther, W. C.: Analysis of variance. Prentice-Hall, Inc., Englewood Cliffs, N. J. (1964).
18. Jackson, R. W. B.: The reliability of mental tests. British J. Psych 29 (1939), 267–287.
19. Mahalanobis, P. C.: Indian Journal of Agricultural Science 2 (1932), 694. (Reference from Snedecor, G.: Statistical methods. The Iowa State College Press, Ames Iowa, U.S.A. (1956)).
20. — Professor Ronald Aylmer Fisher. Sankhya 4 (1938), 265–272. Reprinted in Biometrics 20 (1964), 238–250.

21. Ostle, B.: Statistics in research. The Iowa State University Press, Ames Iowa, U.S.A. (1963).
22. Plackett, R. L.: Models in the analysis of variance. Roy. Stat. Soc. B. 22 (1960), 195–209.
23. Scheffé, H.: Alternative models for the analysis of variance. Ann. Math. Stat. 27 (1956), 251–271.
24. — The analysis of variance. Wiley, New York (1959).
25. Snedecor, G.: Calculation and interpretation of analysis of variance and covariance. Collegiate Press, Inc., Ames Iowa, U.S.A. (1934).
26. "Student": On testing varieties of cereals. Biometrika 15 (1923), 271–293.
27. Sukhatme, P. V.: Sampling theory of surveys with applications. Iowa State College Press, Ames, Iowa (1954).
28. Tippett, L. H. C.: The methods of statistics. Williams and Norgate Ltd., London (1931).
29. Tukey, J. W.: Interaction in a row-by-column design. Mem. Rep. 18, Statist. Res. Group, Princeton University (1949).
30. Wilk, M. B. and O. Kempthorne: Fixed, mixed and random models. J. Amer. Stat. Ass. 50 (1955), 1144–1167.
31. Wold, H.: Causality and Econometrics. Econometrica 22 (1954), 162–177.
32. — Time as the Realm of Forecasting. Symposium on Interdisciplinary Perspectives of Time sponsored by The New York Academy of Sciences, January 17–20 (1966).
33. Yates, F.: Sir Ronald Fisher and the design of experiments. Biometrics 20 (1964), 307–321.

Part II

GENERAL PROBLEMS OF ANALYSIS

CHAPTER 1

THE CHOICE OF MODEL

1.1. Introduction

Using analysis of variance it is very important to build up the correct model. Often in statistical investigations there is, for instance, a need for the variance of some "statistic", say the total mean. With the help of the model equation it is then possible to derive that variance correctly and also to compute the expected mean squares in order to see how the different components in the variance will be estimated.

It is probably still most common to write the model as has been done in part I paragraph 1.1. One writes, for example,

$$y_{ijk} = \mu + a_i + b_j + c_{ij} + e_{ijk}$$

and assumes that a_i, b_j, c_{ij}, and e_{ijk}, are normally and independently distributed random variables. Such a model is constructed in order to satisfy the mathematical requirements of the method of analysis but provides no deeper understanding of the application.

During the 1950's, however, the opinion began to appear in journals that the mathematical models used in analysis of variance should be less mathematical and more closely adapted to the experiment or to the sampling situation. See for example Wilk-Kempthorne (35, 36, 37), who have considered different experimental designs and Cornfield-Tukey (8), who considered sampling data (in the meaning of this paper). One of the points these papers have in common is that the models only reflect how the factor levels have been sampled and (in experiments) randomized to experimental units. As far as possible other assumptions are avoided.

The sampling case, completely randomized experiments and randomized block experiments will be treated here. We are talking about the sampling case when there are no proper experimental plots but we are randomly choosing and investigating, for example small areas ("provytor") from a grain field or samples for analysis from different substances.

In the experimental case the factor levels are sampled (possibly all the population levels can be used in the experiment as in Model I) but these are in some way randomized to the experimental plots. This difference

could also be expressed by saying[1] that for the sampling case every combination of factor levels has its own population of plots, that all combinations have a common plot population in the completely randomized experiment and that groups of combinations have common plot populations in the randomized block experiment.

Both for the sampling case and for the experimental case there is an important distinction between "the population model" and "the statistical model". The population model specifies the structure of the population and the statistical model reflects the procedures of sampling and randomization. The idea of constructing these two models seems to go back to Wilk-Kempthorne (35).

1.2. Population models and statistical models

A simple population model with only one factor is used when one measures the same thing with different methods, for example one and the same physical distance or the percentage of some compound per sample from a homogeneous substance. Then the methods constitute the factor and the "true value" for method (level) I is written

$$\mu(I) = \mu + \alpha(I); \quad I = 1, 2 \ldots A \tag{II.1}$$

where $\alpha(I)$ characterizes method I. The following side condition is generally used:

$$\sum_{I=1}^{A} p_I \cdot \alpha(I) = 0$$

in which p_I are weights and the summation range is over all methods in the population. μ is a weighted mean over all methods

$$\mu = \mu(-) = \Sigma_I p_I \mu(I)$$

if $\Sigma_I p_I = 1$ which in the following will be assumed for all weight systems.

Here the population consists of the non-stochastic values $\mu(1)$, $\mu(2)$, ... $\mu(I)$, ... $\mu(A)$, the structures of which are specified by the population model. There may be infinitely many true values (A infinitely large) and their distribution may be normal but nothing is said about that in the model. The special case of this general model in which the classical assumptions of normality and homogeneity of variances are fulfilled is considered in paragraph 3.2.

In part IV we will meet a case where the population model gives the

[1] See Plackett (21).

structure of constant parameters, exactly as it does here, but where these constants are probabilities. The population may, for example, consist of individuals each with different probabilities for different reactions to a certain stimulus. Then we could say that we have constructed a probability model including stochastic elements but their stochastic nature does not manifest itself until we try to observe them. The population still consists of a finite or an infinite set of constants with a certain fixed structure.

If more than one measurement per method is made, another factor, the measurement error, could be introduced but it does not belong to the population model of true values. Two factors are obtained if we have different methods for producing details of a certain dimension, say the diameter, of special importance. The true value of this dimension will probably vary even between details manufactured in the same way. Then the population model will be:

$$\mu(IJ) = \mu + \alpha(I) + \varepsilon(IJ); \quad \begin{matrix} I = 1, 2 \ \ldots \ A \\ J = 1, 2 \ \ldots \ E_I \end{matrix} \tag{II.2}$$

where $\varepsilon(IJ)$ for different J's express the differences between details from method I. We introduce another side condition:

$$\Sigma_J p_{IJ} \cdot \varepsilon(IJ) = 0$$

which gives the weighted mean

$$\mu(I-) = \Sigma_J p_{IJ} \cdot \mu(IJ) = \mu + \alpha(I)$$

for method I. Here we have two factors, "methods" and "details", the latter of which is said to be nested within the former.

If all the methods in the population are studied we have a sort of stratified sample. Complete investigation of the chosen methods gives one-stage-sampling (if the measurement errors can be ignored) while sampling also of details gives two-stage sampling.

Variance components can be defined in the following way:

$$\sigma^2_{E_I} = \sum_J p_{IJ} \varepsilon^2(IJ); \quad \sigma^2_E = \sum_I p_I \cdot \sigma^2_{E_I}; \quad \sigma^2_A = \sum_I p_I \alpha^2(I)$$

or with equal weights and finite populations

$$\sigma^2_E = \frac{1}{A} \sum_I \frac{1}{E_I - 1} \sum_J \varepsilon^2(IJ); \quad \sigma^2_A = \frac{1}{A-1} \sum_I \alpha^2(I).$$

These models are easily extended by introducing more crossed and/or nested factors.

TABLE II: 1. *The number of observations per combination of treatment and plot in a completely randomized design.*

Plot no.	Treatment no. I	II	III
1	0	1	0
2	0	0	1
3	1	0	0
4	0	0	1
5	1	0	0
6	0	1	0

In a completely randomized experiment with one treatment factor the true response for treatment I applied to plot K can be expressed as follows:

$$\mu(IK) = \mu + \alpha(I) + \varepsilon(K) + \eta(IK); \quad \begin{matrix} I = 1, 2 \ldots A \\ K = 1, 2 \ldots P \end{matrix} \qquad \text{(II.3)}$$

Here $\alpha(I)$ = the effect of treatment I, $\varepsilon(K)$ = the effect of plot K, $\eta(IK)$ = the interaction effect between treatment I and plot K.

According to this model it looks like the two factors "treatments" and "plots" were crossed, but as only one treatment can be applied to every plot the design would in that case be considered incomplete as in table II: 1.

Instead we imagine, that for every treatment we have a population of P true values, namely the responses that would arise if the treatment were applied to all plots. Therefore the factor "plots" can be considered as nested within the treatment factor (at least in the statistical model). However, the responses sampled from these populations will have a certain interdependence because the same plot can only be treated once in the same experiment. This means that this case is not the same as the sampling case with one factor nested within another. As we shall see the interdependence will affect the expected mean squares as long as the plot population is finite.

In a randomized block design with one treatment factor the population model can be written

$$\mu(IJK) = \mu + \alpha(I) + \beta(J) + \gamma(IJ) + \varepsilon(JK) + \eta(IJK); \quad \begin{matrix} I = 1, 2 \ldots A \\ J = 1, 2 \ldots B \\ K = 1, 2 \ldots P_J \end{matrix} \qquad \text{(II.4)}$$

Here we have introduced the block effect $\beta(J)$ and the interaction effect $\gamma(IJ)$ between treatments and blocks. Note that plot effect and interaction effect between treatments and plots are defined within blocks.

With the side conditions

$$\sum_K p^P_{JK} \cdot \varepsilon(JK) = \sum_K p^P_{JK} \cdot \eta(IJK) = 0$$

$$\sum_J p^{Bl}_J \cdot \beta(J) = \sum_J p^{Bl}_J \cdot \gamma(IJ) = 0$$

$$\sum_I p^T_I \cdot \alpha(I) = \sum_I p^T_I \cdot \gamma(IJ) = \sum_I p^T_I \cdot \eta(IJK) = 0$$

the weighted mean value over all experimental plots for combination IJ is:

$$\mu(IJ\text{-}) = \sum_K p^P_{JK} \cdot \mu(IJK) = \mu + \alpha(I) + \beta(J) + \gamma(IJ)$$

and in the same way:

$$\mu(I\text{--}) = \sum_J p^{Bl}_J \cdot \sum_K p^P_{JK} \cdot \mu(IJK) = \mu + \alpha(I)$$

$$\mu(\text{-}J\text{-}) = \sum_I p^T_I \cdot \sum_K p^P_{JK} \cdot \mu(IJK) = \mu + \beta(J)$$

$$\mu(\text{-}JK) = \sum_I p^T_I \cdot \mu(IJK) = \mu + \beta(J) + \varepsilon(JK)$$

$$\mu(\text{---}) = \sum_{IJ} p^T_I \cdot p^{Bl}_J \sum_K p^P_{JK} \cdot \mu(IJK) = \mu$$

The most usual in such an experiment is probably that all the weights are equal.

The statistical model will express exactly what is observed. Let us consider a sampling situation like that of the population model (II.2). Observing the response for combination IJ one will obtain

$$y(IJ) = \mu(IJ) + e(IJ)$$

where an observational technical error of $e(IJ)$ is introduced. These errors are considered as uncorrelated random variables with mean values $= 0$ and variances $\sigma^2_{e_{IJ}}$. As a rule it is possible by careful planning to obtain technical errors that are uncorrelated and have zero means. When the means are zero the situation is equivalent to one where there are no systematic observational errors.

The whole population (all $y(IJ)$'s) is generally not observed; only some levels of factor $\mathcal{A}$ (methods in the example above) and for these only some levels of the factor $\mathcal{E}$ (details in the example). Assume that factor $\mathcal{A}$ has A levels in the population, a of which will be sampled by some procedure and that factor $\mathcal{E}$ has E_I levels for level I of factor $\mathcal{A}$ and that e_I of these will be sampled if I is chosen.

If observed values are indicated with small letters we have the following data to construct a model for:

$$y_{ij}; \quad \begin{matrix} i = 1, 2 \ldots a \\ j = 1, 2 \ldots e_i \end{matrix}$$

If the levels are randomly chosen we do not know, for example, which I corresponds to a certain i. In the model the following so called indicator variables[1] are used:

$$a_i(I) = \begin{cases} 1 & \text{if } i \text{ corresponds to } I \\ 0 & \text{otherwise} \end{cases}$$

$$b_{Ij}(IJ) = \begin{cases} 1 & \text{if } Ij \text{ corresponds to } IJ \\ 0 & \text{otherwise} \end{cases}$$

The statistical model is written:

$$y_{ij} = \mu + \sum_{I=1}^{A} a_i(I)\,\alpha(I) + \sum_{IJ} a_i(I)\,b_{Ij}(IJ)\{\varepsilon(IJ) + e(IJ)\}$$

Another way to write it would be the usual

$$y_{ij} = \mu + a_i + e_{ij}$$

but this form tells us less about what constitutes the random variables a_i and e_{ij} and less as well about their joint distribution. The distributions of the indicator variables $a_i(I)$ and $b_{Ij}(IJ)$ are completely determined by the methods of sampling from the populations of levels. Thus the methods of sampling and the measurement errors completely determine the distribution of y_{ij}. Using simple random sampling the following relations hold for $a_i(I)$:

$$E\{a_i(I)\} = \frac{1}{A};$$

$$E\{a_i(I) \cdot a_{i'}(I')\} = \frac{1}{A(A-1)}; \quad i \neq i' \quad I \neq I'$$

and the corresponding for $b_{Ij}(IJ)$.

In a randomized block experiment with one treatment factor we assume that

[1] Introduced by Cornfield (7).

(i) the treatment factor $\mathcal{A}$ has A levels, a of which will be sampled;

(ii) the block factor $\mathcal{B}$ has B levels, b of which will be sampled;

(iii) within block J there are P_J plots, to r_{IJ} of which level I of the treatment factor will be applied if treatment I and block J are included in the sample.

With the help of the indicator variables

$$a_i(I) = \begin{cases} 1 & \text{if } i \text{ corresponds to } I \\ 0 & \text{otherwise} \end{cases}$$

$$b_j(J) = \begin{cases} 1 & \text{if } j \text{ corresponds to } J \\ 0 & \text{otherwise} \end{cases}$$

$$d_{IJf}(JK) = \begin{cases} 1 & \text{if replication no. } f \text{ of treatment } I \text{ within block } J \text{ has been randomized to plot no. } JK \\ 0 & \text{otherwise} \end{cases}$$

the following model equation is obtained:

$$y_{ijf} = \mu + \sum_{I=1}^{A} a_i(I) \cdot \alpha(I) + \sum_{J=1}^{B} b_j(J) \cdot \beta(J) + \sum_{I=1}^{A} \sum_{J=1}^{B} a_i(I)\, b_j(J) \cdot \gamma(IJ)$$

$$+ \sum_{I=1}^{A} \sum_{J=1}^{B} a_i(I)\, b_j(J) \sum_{K=1}^{P_J} d_{IJf}(JK) \{\varepsilon(JK) + \eta(IJK) + e(IJK)\}; \quad \begin{matrix} i = 1, 2 \ldots a \\ j = 1, 2 \ldots b \\ f = 1, 2 \ldots r_{ij} \end{matrix}$$

If the plot populations are infinite one could just as well use the conventional form of the model:

$$y_{ijf} = \mu + a_i + b_j + c_{ij} + e_{ijf}$$

if one remembers how c_{ij}, a_i and b_j are tied together.

1.3. Applications

Example 1. Estimation of average crop yields (30).

If one wants to estimate the average yield of, for example, autumn wheat in a certain district, one can sample some farms (factor $\mathcal{A}$), for these, some fields (factor $\mathcal{B}$) and for the fields, some small plots for investigation (factor $\mathcal{C}$). The number of farms, that is the number of levels of factor $\mathcal{A}$, in the population is denoted by A and the farms are indicated by I. We further assume that there are B_I fields of autumn wheat on farm I and C_{IJ}

plots on field IJ. The true yield per unit area for plot IJK can then be written:

$$\mu(IJK) = \mu + \alpha(I) + \beta(IJ) + \varepsilon(IJK); \quad \begin{matrix} I = 1, 2 \ldots A \\ J = 1, 2 \ldots B_I \\ K = 1, 2 \ldots C_{IJ} \end{matrix}$$

The sizes of the fields growing autumn wheat vary in one and the same farm, and the area of autumn wheat grown per farm varies from farm to farm. The plots can on the other hand be made equally large. Therefore it is suitable to take simple random samples of plots but to sample fields with probabilities proportional to their sizes. The latter is possible if one measures the sizes of the fields for the chosen farms. The autumn wheat areas for individual farms are unknown, however, and the farms have therefore to be chosen with equal probabilities or probabilities proportional to something else that may be correlated with the area of autumn wheat. In this application the probabilities were made proportional to the total area of arable land. The following side conditions were used:

$$\sum_K \varepsilon(IJK) = \sum_J p_{IJ} \cdot \beta(IJ) = \sum_I p_I \cdot \alpha(I) = 0$$

where $p_{IJ} = \dfrac{\text{area of field } IJ}{\text{autumn wheat area for farm } I}$

and p_I are sampling probabilities. Thus we see that

$$\mu(IJ\text{-}) = \mu + \alpha(I) + \beta(J) = \frac{\text{total yield of field } IJ}{\text{total area of field } IJ}$$

$$\mu(I\text{--}) = \mu + \alpha(I) = \frac{\text{total yield of autumn wheat for farm } I}{\text{total autumn wheat area for farm } I}$$

while $\mu(\text{---}) = \Sigma_I p_I \cdot \mu(I\text{--}) = \mu$ as rule is not equal to

$$\mu' = \frac{\text{total yield of autumn wheat in the district}}{\text{total autumn wheat area in the district}}$$

An unbiased estimate of μ will thus be a biased estimate of μ'. The magnitude of the bias will depend on the correlation between autumn wheat area and total area of arable land.

Plots were sampled by first choosing points on the field, and then harvesting circular areas described using these points as centres. Though the points were well spread to avoid overlap the sample may yet be considered as a simple random sample from an infinite population of plots.

The farms were chosen with probabilities p_I and (theoretically) with replacement. The disadvantage caused by the replacement was thought to

be small, as a relatively small fraction of all the farms were sampled. The number of fields per farm may be very small, and therefore only one field per farm was sampled in one part of the investigation, all the fields in another part.

The statistical model is:

$$y_{ijk} = \mu + \sum_{I} a_i(I)\,\alpha(I) + \sum_{IJ} a_i(I)\,b_{Ij}(IJ)\cdot\beta(IJ)$$
$$+ \sum_{IJK} a_i(I)\,b_{Ij}(IJ)\,d_{IJk}(IJK)\,\{\varepsilon(IJK) + e(IJK)\}; \quad \begin{matrix} i = 1, 2 \ldots a \\ j = 1, 2 \ldots b_i \\ k = 1, 2 \ldots c_{ij} \end{matrix}$$

in which the following indicator variables have been used:

$$a_i(I) = \begin{cases} 1 & \text{if farm } i \text{ is the same as farm } I \\ 0 & \text{otherwise} \end{cases}$$

$$b_{Ij}(IJ) = \begin{cases} 1 & \text{if field } Ij \text{ is the same as field } IJ \\ 0 & \text{otherwise} \end{cases}$$

$$d_{IJk}(IJK) = \begin{cases} 1 & \text{if plot } IJk \text{ is the same as plot } IJK \\ 0 & \text{otherwise} \end{cases}$$

For these the following relations hold:

$$E\{a_i(I)\} = p_I$$

$$E\{a_i(I)\cdot a_{i'}(I')\} = p_I p_{I'}; \quad i \neq i'$$

$$E\{b_{Ij}(IJ)\} = p_{IJ}$$

$$E\{b_{Ij}(IJ)\cdot b_{Ij'}(IJ')\} = p_{IJ}\cdot p_{IJ'}; \quad j \neq j'$$

$$E\{d_{IJk}(IJK)\} = \frac{1}{C_{IJ}}\,(1)$$

$$E\{d_{IJk}(IJK)\cdot d_{IJk'}(IJK')\} = \frac{1}{C_{IJ}(C_{IJ}-1)}; \quad k \neq k';\ K \neq K'$$

The variance components are for example:

$$\sigma^2_{ABC} = \sum_{IJ} p_I p_{IJ}\cdot\frac{1}{C_{IJ}-1}\sum_{K}\varepsilon^2(IJK); \quad \sigma^2_e = \sum_{IJ} p_I p_{IJ}\frac{1}{C_{IJ}}\sum_{K}\sigma^2_{e\,IJK};$$

$$\sigma^2_{AB} = \sum_{IJ} p_I p_{IJ}\beta^2(IJ) \quad etc.$$

[1] Later we let $C_{IJ}\to\infty$.

As we will let $C_{IJ} \to \infty$ there is no reason why $\varepsilon(IJK)$ and $e(IJK)$ should be kept separate. At least with balanced samples (that is when $b_i = b$ and $c_{ij} = c$ for all i and j) one can just as well write the model in the classical way:

$$y_{ijk} = \mu + a_i + b_{ij} + e_{ijk}$$

where a_i, b_{ij} and e_{ijk} are uncorrelated, but dependent, random variables with mean values $= 0$. This can be seen from the following:

$$E(a_i) = \left\{\sum_I a_i(I)\,\alpha(I)\right\} = \sum_I p_I \alpha(I) = 0$$

$$\underset{1,2}{E}(b_{ij}) = \underset{1,2}{E}\left\{\sum_I a_i(I) \sum_J b_{Ij}(IJ) \cdot \beta(IJ)\right\} = \underset{1}{E}\left\{\sum_I a_i(I) \sum_J p_{IJ}\beta(IJ)\right\} = 0^{(1)}$$

$$\underset{1,2}{E}(a_i \cdot b_{ij}) = \underset{1,2}{E}\left\{\sum_I a_i(I) \cdot \alpha(I)\right\}\left\{\sum_I a_i(I) \sum_J b_{Ij}(IJ)\,\beta(IJ)\right\}$$

$$= \underset{1,2}{E}\left\{\sum_I a_i(I)\,\alpha(I) \cdot \sum_J b_{Ij}(IJ) \cdot \beta(IJ)\right\} = \underset{1}{E}\left\{\sum_I a_i(I)\,\alpha(I) \cdot \sum_J p_{IJ}\beta(IJ)\right\} = 0 \quad \text{etc.}$$

In the derivations the fact that $a_i(I) \cdot a_i(I') = 0$ when $I \neq I'$ has been used.

The variances are easily computed:

$$D^2(a_i) = E(a_i^2) = E\left\{\sum_I a_i(I)\,\alpha(I)\right\}^2 = E\left\{\sum_I a_i(I)\,\alpha^2(I)\right\} = \sum_I p_I \alpha^2(I) = \sigma_A^2$$

$$D^2(b_{ij}) = E(b_{ij}^2) = E\left\{\sum_I a_i(I) \sum_J b_{Ij}(IJ)\,\beta(IJ)\right\}^2$$

$$= E\left\{\sum_{II'JJ'} a_i(I)\,a_i(I')\,b_{Ij}(IJ) \cdot b_{I'j}(I'J')\,\beta(IJ) \cdot \beta(I'J')\right\}$$

$$= E\left\{\sum_{IJ} a_i(I)\,b_{Ij}(IJ)\,\beta^2(IJ)\right\} = \sigma_{AB}^2$$

$D^2(e_{ijk}) = \sigma_e^2 + \sigma_{ABC}^2$ in the same way as above

Though we allow unequal variances, the symmetry caused by the random sampling gives (at least formally) equal variances. This will not be the case if, for example, $\mathcal{A}$ is fixed and we put $i = I$, which will be done in deriving variances for individual contrasts.

Example 2. Estimation of different error variances in chemical analyses. In a chemical laboratory routine analyses of the percentages of Fe_2O_3 and ZnO in "smältferrit" were made regularly. "Smältferrit" is a powder kept in large vessels, from which small samples were taken for analysis. To get an idea of the magnitude of the different sources of errors one

[1] $\underset{1,2}{E}$ stands for the expectation taken over the first and second stage, that is sampling of farms and of fields within farms.

started with the following list:

1) variation between different methods of analysis
2) variation between different analysts
3) variation between analyses made by the same analyst at different occasions.
4) variation between multiple analyses made by the same analyst at the same occasion.
5) variation between samples taken from different spots in the vessel.

The last-mentioned source of variation has nothing directly to do with the error of analysis, but with the error that can be committed when generalizing from a small number of samples to the whole vessel. The variation between vessels was not considered pertinent to the investigation. How a model equation including all these factors is to be constructed depends on the design. "Methods" and "analysts" are apparently crossed, "occasions" nested within combinations of "methods" and "analysts", and "multiple analyses" nested within "occasions". The factor "spots in the vessel" could be crossed with "methods" and "analysts", or nested within one of them or a combination of both of them. The investigation was in fact made in smaller parts; in the first, only one method and samples that had been homogenized after sampling were used, in the second part one method, one analyst, but samples from different spots in the vessel. Let us here build up a model for the first 4 factors in the way that has been hinted at above. If we indicate analysts by I, methods by J, occasions by IJK and multiple analyses by $IJKL$ we will have the population model:

$$\mu(IJKL) = \mu + \alpha(I) + \beta(J) + \gamma(IJ) + \delta(IJK) + \varepsilon(IJKL); \qquad \begin{matrix} I = 1, 2 \ldots A \\ J = 1, 2 \ldots B \\ K = 1, 2 \ldots C_{IJ}(1) \\ L = 1, 2 \ldots D_{IJK}(1) \end{matrix}$$

Perhaps it would be more correct to count the last factor in the equation as measurement error and remove it from the population model. In the statistical model below we shall replace $\varepsilon(IJKL)$ by $e(IJK)$.

Suitable side conditions are:

$$\sum_K \delta(IJK) = \sum_J \gamma(IJ) = \sum_J \beta(J) = \sum_I \gamma(IJ) = \sum_I \alpha(I) = 0$$

The statistical model will be written:

$$y_{ijkl} = \mu + \sum_{I=1}^{A} a_i(I)\,\alpha(I) + \sum_{J=1}^{C} b_j(J)\,\beta(J) + \sum_{IJ} a_i(I)\,b_j(J)\,\gamma(IJ)$$

(1) Later we will let C_{IJ} and $D_{IJK} \to \infty$.

$$+\sum_{IJ}\sum_{K=1}^{C_{IJ}} a_i(I)\,b_j(J)\,c_{IJk}(IJK)\,\{\delta(IJK)+e(IJK)\}$$

with the usual meanings of the indicator variables $a_i(I)$, $b_j(J)$ and $c_{IJk}(IJK)$. The $e(IJK)$'s are assumed to be uncorrelated random variables with means $=0$ and variances $\sigma^2_{e_{IJK}}$. The other variance components are defined as follows:

$$\sigma^2_A=\frac{1}{A-1}\sum_I\alpha^2(I);\quad \sigma^2_B=\frac{1}{B-1}\sum_J\beta^2(J);\quad \sigma^2_{AB}=\frac{1}{(A-1)(B-1)}\sum_{IJ}\gamma^2(IJ);$$

$$\sigma^2_{ABC_{IJ}}=\frac{1}{C_{IJ}-1}\sum_K\gamma^2(IJK);\quad \sigma^2_{ABC}=\frac{1}{AB}\sum_{IJ}\sigma^2_{ABC_{IJ}};\quad \sigma^2_e=\sum_{IJ}\frac{1}{C_{IJ}}\sum_K\sigma^2_{e_{IJK}}$$

Here factor $\mathcal{A}$ stands for analysts, factor $\mathcal{B}$ for methods, and factor $\mathcal{C}$ for occasions.

Example 3. Comparing washing-machines (26).

To compare the cleaning ability of different machines one and only one type of (white) garment was washed in them. The experiment consisted of a number of trials. The pieces of garment in the different trials were taken from different batches of dirty linen. The garments from such a batch were distributed differently among the machines in different parts of the experiment. Sometimes the garments were divided into groups according to how dirty they were with as many garments per group as there were machines to be compared. Then one piece of garment per group was randomized to every machine. On other occasions the garments were randomized directly to the machines without any grouping. In the former case we have two systems of blocks: a system of trials and a system of groups nested within trials.

If there are more pieces of garment per trial than will be used in the experiment, the grouping is made on sampled garments and will not be reflected in the population model. The model will then be, irrespective of how the randomization is made:

$$\mu(IJK)=\mu+\alpha(I)+\beta(J)+\gamma(IJ)+\varepsilon(JK)+\eta(IJK);\quad \begin{matrix} I=1,2\ldots A \\ J=1,2\ldots B \\ K=1,2\ldots C_J \end{matrix}$$

with the side conditions

$$\sum_K\varepsilon(JK)=\sum_K\eta(IJK)=\sum_J\gamma(IJ)=\sum_J\beta(J)=\sum_I\eta(IJK)=\sum_I\gamma(IJ)=\sum_I\alpha(I)=0$$

As there is no reason to believe that certain machines (or trials or garments) should be more important than others unequal weights are not required.[1] In the model $\alpha(I)$ stands for machines; $\beta(J)$ for trials, $\gamma(IJ)$ for interaction between machines and trials, $\varepsilon(JK)$ for garments within trial, and $\eta(IJK)$ for interaction between machines and garments within trial.

When the randomization is made directly the statistical model will be exactly as that given for randomized blocks in paragraph 1.2. If the pieces of garment have been grouped, the model will have the same appearance, but now $d_{IJf}(JK)=1$ if that garment within trial J which has come to group f and been randomized to machine I, is garment no. K within trial J. The distribution for this indicator variable will depend on K, as, for example, the dirtiest garment in the population inevitably comes to the "dirtiest group", if that particular garment is sampled. This difficulty arises because we have sampled pieces of garment and not groups. One way of avoiding this would be to disregard the grouping in analysing the experiment—hoping that the grouping has already had some advantageous effect by distributing the pieces of garment among the machines in a more "representative" way.

In this application we had $a=A$, while B and all C_J were considered as infinite, and thus the model can be written (ignoring the problem of grouping):

$$y_{IJf}=\mu+\alpha(I)+b_J+c_{IJ}+e_{IJf}$$

where c_{IJ} is tied to $\alpha(I)$ and b_J.

1.4. Conclusions

The construction of a mathematical model is an integral part of the planning of an investigation. One has to determine the relevant factors that may influence the result, which of them are to be studied, and which kept constant or eliminated by randomization. If a new factor is introduced in the population model, a number of constant terms will enter the model equation. To begin with one main effect for each of the levels of the new factor is included. In addition, an interaction term is included for every combination of the levels of the new factor with the levels of every other factor that is crossed with the new one. In sampling situations we use ordinary nested and crossed factors, experiments also include the plot factor which is of a somewhat different nature. To the population model

[1] In other cases this could be reasonable, for example, in agricultural experiments if the blocks are representing unequally large areas.

is then added a number of suitable side conditions so as not to introduce more independent effects than can be estimated from the true values. Assume, for example, that we have two crossed factors with two levels each and that the four true values are $\mu(11)$, $\mu(12)$, $\mu(21)$ and $\mu(22)$. Thus only four of the parameters

$$\mu,\ \alpha(1),\ \alpha(2),\ \beta(1),\ \beta(2),\ \gamma(11),\ \gamma(12),\ \gamma(21),\ \gamma(22) \quad \text{in the model}$$

$$\mu(IJ) = \mu + \alpha(I) + \beta(J) + \gamma(IJ); \quad \begin{matrix} I = 1, 2 \\ J = 1, 2 \end{matrix}$$

are determined from the true values. This is why we introduce the side conditions

$$\sum_J p_J^B \gamma(IJ) = \sum_J p_J^B \beta(J) = \sum_I p_I^A \gamma(IJ) = \sum_I p_I^A \alpha(I) = 0$$

which leave us four independent parameters.

One must of course already have the analysis in mind when constructing the population model, as an example to be able to choose between making a factor nested or crossed.

The statistical model reflects how the sample is drawn from the possible true values. When the populations are infinite, or the factors fixed, it may seem unnecessary to write it in the complicated way that is used in paragraph 1.2. However the conventional way of writing makes it difficult to visualize the interrelation of the effects and their distribution. The complicated model shows directly that the interaction effect in Model II is tied to the factor effects, and thus it would probably be safer to derive variances and E(MS) from it. In a completely hierarchical design with infinite populations it is better to use the shorter way of writing models, as there are no interactions and no finite population corrections.

CHAPTER 2

ESTIMATION OF VARIANCE COMPONENTS

2.1. Introduction

The main steps of analysis in a situation for analysis of variance are

1) construction of model,
2) synthesis of variances,
3) partitioning of the total sum of squares,
4) derivation of expected mean squares,
5) estimation of variance components,
6) testing of hypotheses,
7) estimation of contrasts and their variances.

Steps 6) and 7) will be treated in chapters 3 and 4, the other steps in this chapter. The total sum of squares will sometimes be divided in orthogonal components, sometimes we will define and use mean squares that are far from orthogonal and may even be only small modifications of each other. We shall be mostly concerned with defining suitable mean squares, so that we shall be able to obtain unbiased estimates of the variance components. This means that the expected mean squares will play a central role. The estimates will be obtained by equating the computed mean squares to the expected mean squares, and solving the resulting system of equations. Sometimes this system contains more unknown components than equations but as a rule it is still possible to estimate what we want to estimate, for example the variance of the total mean.

In the following we will distinguish between balanced and unbalanced samples. A sample is said to be balanced if there is an equal number of observations per factor combination and the same number of levels of factor $\mathcal{B}$ for every level of factor $\mathcal{A}$ in case $\mathcal{B}$ is nested within $\mathcal{A}$. The balanced case will be treated in 2.2 the unbalanced in 2.3. In paragraph 2.4 some short comments will be made on the possibility of computing variances and interval estimates for the variance components.

2.2. Balanced samples

2.2.1. *A small introductory example*

In a comparison of the percentage of bleaching-powder in different washing-materials (29) one wanted to give tolerance limits for a standard

dose from every brand. That is, an answer to the following question was wanted: If a house-wife buys a package of washing-material and takes a dose from it, within which limits can she expect the percentage of bleaching-powder to be? In order to obtain this range of percentages it was considered that a number of packages would be bought, and from these, samples, the size of a standard dose, taken out for analysis. The number of packages and samples per package in the population may here be considered as infinite. The statistical model can therefore be written out directly as follows:

$$y_{ij} = \mu + a_i + b_{ij} + e_{ij}$$

where

y_{ij} = observed percentage for sample j within package i,
a_i = effect of package i,
b_{ij} = effect of sample j from package i,
e_{ij} = measurement error,

and a, b and e are uncorrelated random variables with means = 0 and variances σ_A^2, σ_{AB}^2 and σ_e^2. The two latter variances may be averages of different variances for different packages. It is also possible to let the variance of the measurement error vary from sample to sample within packages.

One wants to set tolerance limits in the distribution of y_{ij} and therefore need its variance

$$D^2(y_{ij}) = \sigma_e^2 + \sigma_{AB}^2 + \sigma_A^2$$

This variance will be estimated with the help of analysis of variance. As this will require only the most simple application of the analysis of variance, dividing the total sum of squares in "between" and "within" packages, we get the expected mean squares directly:

$$\mathrm{E}(\mathrm{MS}_A) = \sigma_e^2 + \sigma_{AB}^2 + r\sigma_A^2$$

$$\mathrm{E}(\mathrm{MS}_{AB}) = \sigma_e^2 + \sigma_{AB}^2$$

We have here used the notations MS_A and MS_{AB} for the mean squares in the analysis of variance and r for the number of samples per package. Equally many samples per package will make the sample balanced.

The variance $D^2(y_{ij})$ will now be estimated unbiasedly by

$$\hat{D}^2(y_{ij}) = \frac{\mathrm{MS}_A + (r-1)\,\mathrm{MS}_{AB}}{r}$$

2.2.2. *Expected mean squares*

When the sample is balanced the usual anova-mean-squares[1] are always used. How these will be computed is immediately clear at least in simpler cases. Rules for deriving them in an arbitrary case are given, for example, by Scheffé (24).

To be able to make estimates and tests of hypotheses it is necessary to know the expected mean squares. Discussions of how they are to be computed and rules for how they can be written out without greater effort can be found in many journals and text books. Rules for Model I, Model II and mixed model in the sampling case are given, for example, by Schultz (25) and Scheffé (24). Scheffé's method is a special case of one given by Bennett–Franklin (3), who consider finite populations. The case of finite populations is also treated by Cornfield–Tukey (8). These references give rules for how the E(MS)'s are to be constructed. In a report by Abt (1) they are written out for 2, 3, and 4 crossed factors with the three classical models (no finite non-exhausted populations). Thus the sampling case is well covered by rules; what is lacking, at least in the above mentioned references, are methods for unbalanced populations. However, Cornfield–Tukey (8) allow unequal populations for the last stage in the sampling. Unbalanced populations are not too unusual; see, for example, the investigation of average yield in example 1 above, where there were an unequal number of fields per farm. Perhaps still more common is that an unbalanced population is combined with an unbalanced sample.

In experiments, on the other hand, it is not so easy to write out E(MS) for an arbitrary case directly. Wilk–Kempthorne (37) have given expected mean squares for up to 2 crossed factors in completely randomized designs and up to 2 crossed treatment factors without unit-treatment-interaction in randomized block designs. Zyskind (38) has used certain linear combinations of variance components, cap sigmas (Σ) introduced by Wilk (33), to give rules as general as in the sampling case. These rules, however, use so many notations and are so complicated that they are not very useful in practice.

As has been pointed out before there is no difference between the sampling case and the experimental case when the plot population(s) is (are) infinite.

2.2.3. *The sampling case*

For this case the Cornfield–Tukey's rules are used and then it is necessary to identify the components into which the total mean square will be divided

[1] anova = *an*alysis of *va*riance.

TABLE II: 2. *Plan over the factor combinations when factor $\mathcal{B}$ with b levels is nested within factor $\mathcal{A}$ with a levels, and factor $\mathcal{C}$ with c levels is crossed with $\mathcal{B}$. Factor $\mathcal{E}$ with e levels is nested within $\mathcal{ABC}$, that is, it has e different levels for each combination of the factors $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$.*

Factor $\mathcal{A}$, level no.		1	2	...	a
Factor $\mathcal{B}$, level no.		1 2 ... b	1 2 ... b	...	1 2 ... b
Factor $\mathcal{C}$ level no.	1	1 1 ... 1 2 2 ... 2 ⋮ ⋮ e e ... e			
	2	⋮			
	⋮	⋮			
	c	⋮			

and denote them suitably. If, for example, factor $\mathcal{B}$ is nested within $\mathcal{A}$, $\mathcal{C}$ crossed with $\mathcal{B}$, and $\mathcal{E}$ nested within $\mathcal{ABC}$ the plan in table II: 2 over the $a \cdot b \cdot c \cdot d$ observations is obtained.

The total sum of squares is now divided into the components denoted by

A: effect of factor $\mathcal{A}$,
AB: effect of factor $\mathcal{B}$ within $\mathcal{A}$,
C: effect of factor $\mathcal{C}$,
AC: interaction between $\mathcal{A}$ and $\mathcal{C}$,
ABC: interaction between $\mathcal{B}$ and $\mathcal{C}$ within $\mathcal{A}$,
ABCE: effect of factor $\mathcal{E}$ within $\mathcal{ABC}$.

In these notations two types of letters are distinguished, on the one hand letters specifying the main effect or interaction, on the other hand letters that specify the factor or factor combination in which the main effect or interaction is nested. The former are by Cornfield–Tukey denoted "interactionlike", the latter "not interactionlike". Other names are "essential" and "nonessential" (Schultz), "live" and "dead" (Scheffé). In the example above, for the component AB, B is interactionlike but not A, for the component AC both A and C, and for the component $ABCE$ only E etc.

Each of these components corresponds to both one variance component and one expected mean square. For every variance component it is now easy to determine its coefficient in any expected mean square. The letter Q contributes to that coefficient according to table II: 3.

TABLE II: 3. *Auxiliary table for constructing expected mean squares in the sampling case. The factor Q has Q levels in the population and q in the sample.*

Subscripts of mean squares	Subscripts of variance components: Q present and interactionlike	Q present and not interactionlike	Q absent
Q present	1	1	0
Q absent	$1-q/Q$	1	q

Let us consider $E(MS_A)$. The following components will be obtained:

Component	Coefficient with contribution from A	B	C	E
σ^2_A	1	b	c	e
σ^2_{AB}	1	$\left(1-\frac{b}{B}\right)$	c	e
σ^2_C	0			
σ^2_{AC}	1	b	$\left(1-\frac{c}{C}\right)$	e
σ^2_{ABC}	1	$\left(1-\frac{b}{B}\right)$	$\left(1-\frac{c}{C}\right)$	e
σ^2_{ABCE}	1	1	1	$\left(1-\frac{e}{E}\right)$

This gives: $E(MS_A)=\left(1-\frac{e}{E}\right)\sigma^2_{ABCE}+e\left(1-\frac{c}{C}\right)\left(1-\frac{b}{B}\right)\sigma^2_{ABC}+eb\left(1-\frac{c}{C}\right)\sigma^2_{AC}$

$+ec\left(1-\frac{b}{B}\right)\sigma^2_{AB}+ecb\,\sigma^2_A.$

If one introduces a factor $\mathcal{M}$ of measurement errors with an infinite population of possible errors for every combination $ABCE$, its average variance σ^2_{ABCEM} or σ^2_e will have the coefficient 1 in every E(MS).

The population model of observations for this example would be

$$y(IJKL)=\mu+\alpha(I)+\beta(IJ)+\gamma(K)+\delta(IK)+\eta(IJK)$$
$$+\varepsilon(IJKL)+e(IJKL);\quad I=1,2\ldots A\quad K=1,2\ldots C$$
$$J=1,2\ldots B\quad L=1,2\ldots E$$

with obvious notations and side conditions and the following variance components:

$$\sigma_A^2 = \frac{1}{A-1}\sum_{I=1}^{A}\alpha^2(I); \quad \sigma_{AB}^2 = \frac{1}{A(B-1)}\sum_{I=1}^{A}\sum_{J=1}^{B}\beta^2(IJ);$$

$$\sigma_C^2 = \frac{1}{C-1}\sum_{K=1}^{C}\gamma^2(K); \quad \sigma_{AC}^2 = \frac{1}{(A-1)(C-1)}\sum_{IK}\delta^2(IK);$$

$$\sigma_{ABC}^2 = \frac{1}{A(B-1)(C-1)}\sum_{IJK}\eta^2(IJK); \quad \sigma_{ABCE}^2 = \frac{1}{ABC}\cdot\frac{1}{E-1}\sum_{IJKL}\varepsilon^2(IJKL)$$

$$= \frac{1}{ABC}\sum_{IJK}\sigma_{IJK}^2; \quad \sigma_e^2 = \frac{1}{ABCE}\sum_{IJKL}\sigma_{e_{IJKL}}^2$$

Cornfield–Tukey considered in detail only the case of two crossed factors and for that case they introduced a population as unbalanced as possible. Using our present example and notation, the combination ABC has E_{IJK} levels for factor $\mathcal{E}$. The E(MS) will then include

$$\frac{1}{ABC}\sum_{IJK}\left(1-\frac{e}{E_{IJK}}\right)\sigma_{IJK}^2$$

which Cornfield–Tukey put equal to $(1-e/E)\sigma_{ABCE}^2$ with unchanged E(MS)'s as the result. This new E they called the "effective population size".

If we introduce an unbalanced population, $J = 1, 2 \ldots B_I$, it is easy to see that $(1-b/B)\sigma_{AB}^2$ and $(1-b/B)\sigma_{ABC}^2$ in the balanced case will be replaced by $(1/A)\,\Sigma_I(1-b/B_I)\sigma_{AB_I}^2$ and $(1/A)\,\Sigma_I(1-b/B_I)\sigma_{ABCI}^2$ respectively in the unbalanced. This is in complete analogy with Cornfield–Tukey's result and the E(MS)'s may be left unchanged if a less cumbersome expression is desired. In that case an effective population size B has been introduced.

Often the aim of the analysis is to estimate the variance for the observed total mean. This variance is obtained from Cornfield–Tukey's rules by deriving the expectation for a mean square without subscripts and dividing it with the total number of observations. For the above example the following is obtained:

$$D^2(y.....) = \frac{1}{abce}\left[\left(1-\frac{e}{E}\right)\sigma_{ABCE}^2 + e\left(1-\frac{c}{C}\right)\left(1-\frac{b}{B}\right)\sigma_{ABC}^2\right.$$

$$\left. + eb\left(1-\frac{c}{C}\right)\left(1-\frac{a}{A}\right)\sigma_{AC}^2 + eba\left(1-\frac{c}{C}\right)\sigma_C^2 + ec\left(1-\frac{b}{B}\right)\sigma_{AB}^2 + ecb\left(1-\frac{a}{A}\right)\sigma_A^2\right]$$

2.2.4. *The experimental case*

Complete randomization.—The main difference between the sampling case and the experimental case lies in the experimental plots. In both

cases we sample some factor combinations, which in experiments are randomized to plots; whereas they are investigated directly in sampling situations. For completely randomized designs we assume that there is a population of P plots, from which, in the above example, $a \cdot b \cdot c \cdot e \cdot r$ are sampled and randomized with r to each of the combinations $ABCE$. To the equation for the population model will now be added the plot effect $\pi(M)$ and the unit-treatment-interaction effect $\varphi(IJKLM)$. However, the latter has to be split up into separate interactions between individual factors and plots:

$$\varphi(IJKLM) = \alpha\varphi(IM) + \beta\varphi(IJM) + \gamma\varphi(KM) + \delta\varphi(IKM) + \eta\varphi(IJKM) + \varepsilon\varphi(IJKLM)$$

The usual side conditions for these components are that they are $=0$ if summed over an interactionlike letter, for instance:

$$\sum_J \beta\varphi(IJM) = \sum_M \beta\varphi(IJM) = 0$$

where J and M specify an interaction nested within a factor specified by I.

Rules will now be given for the construction of expected mean squares in completely randomized designs with (to begin with) balanced populations. At first we define the variance components for the interaction between each treatment factor and plots as follows:

$$\sigma^2_{AP} = \frac{1}{A(P-1)} \sum_{IM} (\alpha\varphi)^2\{IM\}; \quad \sigma^2_{ABP} = \frac{1}{AB(P-1)} \sum_{IJM} (\beta\varphi)^2\{IJM\};$$

$$\sigma^2_{ABCEP} = \frac{1}{ABCE(P-1)} \sum_{IJKLM} (\varepsilon\varphi)^2\{IJKLM\} \text{ etc.}$$

Note that only the number of plots has been diminished by 1 in the denominators. In addition we introduce a variance component for plots and a pooled component for technical errors:

$$\sigma^2_P = \frac{1}{P-1} \sum_M \pi^2(M); \quad \sigma^2_e = \frac{1}{ABCEP} \sum_{IJKLM} \sigma^2_{e_{IJKLM}}$$

The table of E(MS) is now obtained from the corresponding table derived by Cornfield–Tukey's rules in the following way:

1) To every E(MS) is added $\sigma^2 = \sigma^2_e + \sigma^2_P + \sigma^2_{AP} + \ldots + \sigma^2_{ABCEP}$, that is the sum of the variance components for technical errors, plots and all the interactions between treatments and plots.

2) The other components are replaced by the component minus $1/P \cdot$ constant $\cdot$ corresponding interaction between component and plots. If the letter Q is interactionlike among the letters characterizing the variance component the factor $Q/(Q-1)$ is included in the constant.
3) A within-mean-square, MS_W, will be added if the number r of plots per treatment combination >1. The expectation is $E(MS_W)=\sigma^2$.

As an example consider $E(MS_A)$, derived for the sampling case on page 47. The result is here:

$$E(MS_A)=\sigma^2+r\left(1-\frac{e}{E}\right)\left[\sigma^2_{ABCE}-\frac{1}{P}\cdot\frac{E}{E-1}\sigma^2_{ABCEP}\right]$$

$$+re\left(1-\frac{c}{C}\right)\left(1-\frac{b}{B}\right)\left[\sigma^2_{ABC}-\frac{1}{P}\cdot\frac{B}{B-1}\cdot\frac{C}{C-1}\cdot\sigma^2_{ABCP}\right]$$

$$+reb\left(1-\frac{c}{C}\right)\left[\sigma^2_{AC}-\frac{1}{P}\cdot\frac{A}{A-1}\cdot\frac{C}{C-1}\cdot\sigma^2_{ACP}\right]$$

$$+rec\left(1-\frac{b}{B}\right)\left[\sigma^2_{AB}-\frac{1}{P}\cdot\frac{B}{B-1}\cdot\sigma^2_{ABP}\right]+recb\left[\sigma^2_{A}-\frac{1}{P}\cdot\frac{A}{A-1}\cdot\sigma^2_{AP}\right]$$

In treating the case with two crossed factors Wilk–Kempthorne (35, 37) defined some of the variance components differently. In our notations these components correspond to the following:

$$Q^2_{ABCEP}=\frac{E}{E-1}\sigma^2_{ABCEP};\quad Q^2_{ABCP}=\frac{B}{B-1}\frac{C}{C-1}\sigma^2_{ABCP}\quad\ldots$$

$$\sigma^2_{\varphi}=\sigma^2_{AP}+\sigma^2_{ABP}+\ldots+\sigma^2_{ABCEP}$$

Thus they have used $Q-1$ instead of Q in the denominators of the unit-treatment-interactions if Q is interactionlike and then they have added all these interactions to a new component σ^2_{φ}. In this way they obtain somewhat shorter expressions but on the other hand one variance component more than in the present paper.

The rules can as before even be applied to the variance for the total mean, $D^2(y\ldots..)$.

As another example we consider the case with two factors $\mathcal{A}$ and $\mathcal{B}$ of which $\mathcal{B}$ is nested within $\mathcal{A}$. For this example it is relatively easy to go through the algebra and thus check the rules. The population model can be written as:

$$\mu(IJK) = \mu + \alpha(I) + \beta(IJ) + \varepsilon(K) + \eta_1(IK) + \eta_2(IJK); \quad \begin{matrix} I = 1, 2 \ldots A \\ J = 1, 2 \ldots B \\ K = 1, 2 \ldots P \end{matrix}$$

$$\alpha(-) = \beta(I-) = \varepsilon(-) = \eta_1(I-) = \eta_1(-K) = \eta_2(IJ-) = \eta_2(I-K) = 0.^1$$

Here $\varepsilon(K)$ is the effect of plot K, $\eta_1(IK)$ the interaction effect between factor $\mathcal{A}$ and plots, and $\eta_2(IJK)$ the interaction effect between factor $\mathcal{B}$ and plots within factor $\mathcal{A}$.

If the observations are

$$y_{ijf}; \quad \begin{matrix} i = 1, 2 \ldots a \\ j = 1, 2 \ldots b \\ f = 1, 2 \ldots r \end{matrix}$$

we obtain the following expected mean squares:

$$\mathrm{E(MS}_A) = \sigma^2 + r\left(1 - \frac{b}{B}\right)\left[\sigma^2_{AB} - \frac{1}{P} \cdot \frac{B}{B-1}\sigma^2_{ABP}\right] + rb\left[\sigma^2_A - \frac{1}{P} \cdot \frac{A}{A-1} \cdot \sigma^2_{AP}\right]$$

$$\mathrm{E(MS}_{AB}) = \sigma^2 + r\left[\sigma^2_{AB} - \frac{1}{P} \cdot \frac{B}{B-1} \cdot \sigma^2_{ABP}\right]$$

$$\mathrm{E(MS}_W) = \sigma^2 = \sigma^2_e + \sigma^2_P + \sigma^2_{AP} + \sigma^2_{ABP}$$

When the population is unbalanced, $J = 1, 2 \ldots B_I$, it is easy to see that the following modifications are obtained:

$$\mathrm{E(MS}_A) = \sigma^2 + r\left[\frac{1}{A}\sum_I \left(1 - \frac{b}{B_I}\right)\sigma^2_{AB_I} - \frac{1}{P} \cdot \frac{1}{A}\sum_I \left(1 - \frac{b}{B_I}\right) \cdot \frac{B_I}{B_I - 1}\sigma^2_{ABP_I}\right]$$
$$+ rb\left[\sigma^2_A - \frac{1}{P} \cdot \frac{A}{A-1} \cdot \sigma^2_{AP}\right]$$

$$\mathrm{E(MS}_{AB}) = \sigma^2 + r\left[\sigma^2_{AB} - \frac{1}{P}\frac{1}{A}\sum_I \frac{B_I}{B_I - 1}\sigma^2_{ABP_I}\right]$$

As in the sampling case it would be possible to introduce:

$$\frac{1}{A}\sum_I \left(1 - \frac{b}{B_I}\right)\sigma^2_{AB_I} = \left(1 - \frac{b}{B'}\right)\sigma^2_{AB} \quad \text{and}$$

$$\frac{1}{A}\sum_I \left(1 - \frac{b}{B_I}\right)\frac{B_I}{B_I - 1}\sigma^2_{ABP_I} = \left(1 - \frac{b}{B''}\right)\frac{1}{A}\sum_I \frac{B_I}{B_I - 1}\sigma^2_{ABP_I}$$

but the effective population sizes B' and B'' will as a rule not be equal.

[1] When an index is replaced by a – the mean over the population has been taken.

It can be mentioned that the checking of formulas for E(MS) can be very much simplified. Only unit-treatment-interactions need to be considered, everything else follows directly from Cornfield–Tukey's rules.

Randomized blocks.—The case of one treatment-factor has been considered by Wilk–Kempthorne (35, 37) and Seeger (26, 28), the results of which differ only in notation. Here we will use the following population model

$$\mu(IJK) = \mu + \alpha(I) + \beta(J) + \gamma(IJ) + \varepsilon(JK) + \eta(IJK); \quad \begin{matrix} I = 1, 2 \ldots A \\ J = 1, 2 \ldots B \\ K = 1, 2 \ldots P \end{matrix}$$

where

$\alpha(I)$ = effect of treatment-factor $\mathcal{A}$'s level I,
$\beta(J)$ = effect of block-factor $\mathcal{B}$'s level J,
$\gamma(IJ)$ = block-treatment interaction,
$\varepsilon(JK)$ = effect of plot K within block J,
$\eta(IJK)$ = unit-treatment interaction within block J

obey the following conditions

$$\alpha(-) = \beta(-) = \gamma(I-) = \gamma(-J) = \varepsilon(J-) = \eta(IJ-) = \eta(-JK) = 0$$

The variance components will be defined as in (26) but with notations changed so as to make it convenient to use the rules of Cornfield–Tukey:

$$\sigma_A^2 = \frac{1}{A-1}\sum_I \alpha^2(I); \quad \sigma_B^2 = \frac{1}{B-1}\sum_J \beta^2(J); \quad \sigma_{AB}^2 = \frac{1}{(A-1)(B-1)}\sum_{IJ} \gamma^2(IJ);$$

$$\sigma_{BP}^2 = \frac{1}{B(P-1)}\sum_{JK} \varepsilon^2(JK); \quad \sigma_{ABP}^2 = \frac{1}{AB(P-1)}\sum_{IJK} \eta^2(IJK)$$

The observed values are

$$y_{ijf}; \quad \begin{matrix} i = 1, 2 \ldots a \\ j = 1, 2 \ldots b \\ f = 1, 2 \ldots r \end{matrix}$$

Using these notations and rearranging the results in (26) one obtains the following expected mean squares:

$$\mathrm{E(MS}_A) = \sigma^2 + r\left(1 - \frac{b}{B}\right)\left[\sigma_{AB}^2 - \frac{1}{P}\cdot\frac{A}{A-1}\cdot\sigma_{ABP}^2\right] + rb\left[\sigma_A^2 - \frac{1}{BP}\cdot\frac{A}{A-1}\cdot\sigma_{ABP}^2\right]$$

$$\mathrm{E(MS}_B) = \sigma^2 + r\left(1 - \frac{a}{A}\right)\left[\sigma_{AB}^2 - \frac{1}{P}\cdot\frac{A}{A-1}\cdot\sigma_{ABP}^2\right] + ra\left[\sigma_B^2 - \frac{1}{P}\cdot\sigma_{BP}^2\right]$$

$$\mathrm{E(MS}_{AB}) = \sigma^2 + r\left[\sigma_{AB}^2 - \frac{1}{P}\cdot\frac{A}{A-1}\cdot\sigma_{ABP}^2\right]$$

$$\mathrm{E(MS}_{ABP}) = \sigma^2 = \sigma_\varepsilon^2 + \sigma_{BP}^2 + \sigma_{ABP}^2$$

This leads us to modify the rules for writing out E(MS) given in the completely randomized case as follows:

If we always denote the block-factor by $\mathcal{B}$, the number of blocks in the population by B and the number of plots per block by P we can replace rule 2) on page 50 by:

$2)^{RB}$ A variance component $\sigma^2_{Q_1 Q_2 \dots Q_s}$, where $Q_1 \dots Q_s$ stand for treatment-factors, will be changed to

$$\sigma^2_{Q_1 Q_2 \cdots Q_s} - \frac{1}{BP} \pi \frac{Q_i}{Q_i - 1} \cdot \sigma^2_{Q_1 Q_2 \dots Q_s BP}$$

and $\sigma^2_{Q_1 Q_2 \dots Q_s B}$

to $$\sigma^2_{Q_1 Q_2 \dots Q_s B} - \frac{1}{P} \pi \frac{Q_i}{Q_i - 1} \sigma^2_{Q_1 Q_2 \dots Q_s BP}$$

The product $\pi(Q_i/Q_i - 1)$ is taken over all interactionlike indices among the $Q_1 \dots Q_s$.

It is now easy to write out E(MS) for any structure of treatments. Let us do so for two crossed factors $\mathcal{A}$ and $\mathcal{C}$. The result is shown in table II: 4.

As in completely randomized designs, it is possible to introduce unbalanced populations at the price of still more complicated expressions.

2.2.5. *Discussion*

In this paragraph rules are given for how to construct tables of expected mean squares in the sampling case and the completely randomized and the randomized block experiment. If the fractions a/A; b/B etc. are known, it is possible to estimate certain variance components. In the sampling case the ordinary components σ^2_A; σ^2_{AB} etc. are estimated, in the experimental case new components of the type $[\sigma^2_A - 1/P \cdot A/(A-1) \cdot \sigma^2_{AP}]$. These components enter the variances, that we may want to synthesize, for example $D^2(y \dots)$, which also can be derived by the rules. In chapter 4 it will be illustrated how the variances for different contrasts in means can be estimated.

As can be seen from the rules, the only difference between the sampling and experimental cases when the plot populations are infinite is that σ^2 in the experimental case includes not only technical errors, σ^2_e, but also plot-errors and unit-treatment-interactions. Thus with infinite populations of plots there is only need for the simpler rules in the sampling case. The

TABLE II: 4. *Expected mean squares in a balanced case of randomized blocks (factor B) with two crossed treatment factors A and C. Each combination is replicated r times in every block.*

$$E(MS_A) = \sigma^2 + r\left(1-\frac{b}{B}\right)\left(1-\frac{c}{C}\right)\left[\sigma^2_{ACB} - \frac{1}{P}\cdot\frac{A}{A-1}\cdot\frac{C}{C-1}\cdot\sigma^2_{ACBP}\right]$$
$$+ rb\left(1-\frac{c}{C}\right)\left[\sigma^2_{AC} - \frac{1}{BP}\cdot\frac{A}{A-1}\cdot\frac{C}{C-1}\cdot\sigma^2_{ACBP}\right]$$
$$+ rc\left(1-\frac{b}{B}\right)\left[\sigma^2_{AB} - \frac{1}{P}\cdot\frac{A}{A-1}\cdot\sigma^2_{ABP}\right] + rbc\left[\sigma^2_A - \frac{1}{BP}\cdot\frac{A}{A-1}\cdot\sigma^2_{ABP}\right]$$

$$E(MS_C) = \sigma^2 + r\left(1-\frac{b}{B}\right)\left(1-\frac{c}{C}\right)\left[\sigma^2_{ACB} - \frac{1}{P}\cdot\frac{A}{A-1}\cdot\frac{C}{C-1}\cdot\sigma^2_{ACBP}\right]$$
$$+ rb\left(1-\frac{a}{A}\right)\left[\sigma^2_{AC} - \frac{1}{BP}\cdot\frac{A}{A-1}\cdot\frac{C}{C-1}\cdot\sigma^2_{ACBP}\right]$$
$$+ ra\left(1-\frac{b}{B}\right)\left[\sigma^2_{CB} - \frac{1}{P}\cdot\frac{C}{C-1}\cdot\sigma^2_{CBP}\right] + rba\left[\sigma^2_C - \frac{1}{BP}\cdot\frac{C}{C-1}\cdot\sigma^2_{CBP}\right]$$

$$E(MS_{AC}) = \sigma^2 + r\left(1-\frac{b}{B}\right)\left[\sigma^2_{ACB} - \frac{1}{P}\cdot\frac{A}{A-1}\cdot\frac{C}{C-1}\cdot\sigma^2_{ACBP}\right]$$
$$+ rb\left[\sigma^2_{AC} - \frac{1}{BP}\cdot\frac{A}{A-1}\cdot\frac{C}{C-1}\cdot\sigma^2_{ACBP}\right]$$

$$E(MS_B) = \sigma^2 + r\left(1-\frac{a}{A}\right)\left(1-\frac{c}{C}\right)\left[\sigma^2_{ACB} - \frac{1}{P}\cdot\frac{A}{A-1}\cdot\frac{C}{C-1}\cdot\sigma^2_{ACBP}\right]$$
$$+ rc\left(1-\frac{a}{A}\right)\left[\sigma^2_{AB} - \frac{1}{P}\cdot\frac{A}{A-1}\cdot\sigma^2_{ABP}\right]$$
$$+ ra\left(1-\frac{c}{C}\right)\left[\sigma^2_{CB} - \frac{1}{P}\cdot\frac{C}{C-1}\cdot\sigma^2_{CBP}\right] + rac\left[\sigma^2_B - \frac{1}{P}\cdot\sigma^2_{BP}\right]$$

$$E(MS_{AB}) = \sigma^2 + r\left(1-\frac{c}{C}\right)\left[\sigma^2_{ACB} - \frac{1}{P}\cdot\frac{A}{A-1}\cdot\frac{C}{C-1}\cdot\sigma^2_{ACBP}\right]$$
$$+ rc\left[\sigma^2_{AB} - \frac{1}{P}\cdot\frac{A}{A-1}\cdot\sigma^2_{ABP}\right]$$

$$E(MS_{CB}) = \sigma^2 + r\left(1-\frac{a}{A}\right)\left[\sigma^2_{ACB} - \frac{1}{P}\cdot\frac{A}{A-1}\cdot\frac{C}{C-1}\cdot\sigma^2_{ACBP}\right] + ra\left[\sigma^2_{CB} - \frac{1}{P}\cdot\frac{C}{C-1}\cdot\sigma^2_{CBP}\right]$$

$$E(MS_{ACB}) = \sigma^2 + r\left[\sigma^2_{ACB} - \frac{1}{P}\cdot\frac{A}{A-1}\cdot\frac{C}{C-1}\cdot\sigma^2_{ACBP}\right]$$

$$E(MS_{ACBP}) = \sigma^2 = \sigma^2_e + \sigma^2_{BP} + \sigma^2_{ABP} + \sigma^2_{CBP} + \sigma^2_{ACBP}$$

---

question is then: do we ever encounter finite plot populations in practice, that is do we need the rules in the experimental case? For instance, even if all the plots in a field (or only a sample of them) are being used, we may want to draw the inferences about some other, vague but infinite, population of equivalent plots. In one example discussed above the blocks consisted of different "batches" of garments each with a fictitious infinite population of garments. However, it is possible to have blocks where all possible plots in the blocks are used, for example, when blocks are days and the plots the three shifts. In that case the number of plots per block is always $P=3$.

Thus there is sometimes a need in applications for the more complicated rules, and besides, they will contribute to a better understanding of the models.

2.2.6. *Applications*

Example 2. Estimation of different error variances in chemical analyses (page 38).

From the situation described in paragraph 1.3 we take out the case in which 3 analysts made double-analyses of Fe_2O_3-percentages on 8 occasions each (homogeneous samples). This is apparently a sampling case and with an infinite number of potential occasions and analyses per occasion. The 3 analysts were considered as representative of an infinite population of analysts that could be employed at the laboratory. Thus the populations are infinite and we can write the statistical model as

$$y_{ijf} = \mu + a_i + b_{ij} + e_{ijf}$$

but the more complicated expression used in 1.3 would be more informative. It could for example show that the variances between occasions may be

TABLE II: 5. *Results in % of Fe_2O_3 of chemical analyses of "smältferrit".*

Date for occasion	Analyst no. 1		2		3	
21/11					70.39	70.21
22/11					70.29	70.36
29/11					70.31	70.38
30/11	70.34	70.56	70.26	70.26	70.31	70.40
2/12	70.33	70.34	70.27	70.31		
5/12			70.33	70.27		
6/12	70.32	70.26			70.18	70.34
8/12	70.30	70.32			70.29	70.29
9/12			70.27	70.28		
14/12					70.29	70.27
15/12	70.35	70.31	70.27	70.30		
16/12	70.36	70.47			70.28	70.31
17/12			70.26	70.28		
19/12	70.43	70.30	70.29	70.32		
20/12	70.17	70.31				
22/12			70.23	70.27		

different for different analysts, and that the variances within occasions may differ.

Occasions must be considered as nested within analysts, especially as occasion j as a rule occurred on different dates for different analysts. The primary values are shown in table II: 5.

The analysis gives the following table:

Sources of variation	D.f.	MS	E(MS)
Analysts	2	0.015728	$\sigma^2 + 2\sigma^2_{AB} + 16\sigma^2_A$
Occasions within analysts	21	0.003991	$\sigma^2 + 2\sigma^2_{AB}$
Within occasions	24	0.003938	σ^2

From this an estimate of the standard deviation is obtained
between analysts to $\hat{\sigma}_A = 0.03$,
between occasions within analysts to $\hat{\sigma}_{AB} = 0.005$,
within occasions to $\hat{\sigma} = 0.06$.

A separate analysis for each analyst gave

$$\hat{\sigma}_1 = 0.08; \quad \hat{\sigma}_2 = 0.02; \quad \hat{\sigma}_3 = 0.07$$

which indicate that each analyst gives rise to errors of different magnitudes and thus that the average σ may not be so informative.

Example 4. Estimation of variance in yield per "hektar" between districts and years.

As a basis for computing the franchise insurance (the level of value of crops destroyed below which the farmer must himself defray the costs involved) etc. in the Swedish crop-insurance system two types of data on yields were used. One type consisted of subjective estimations made by so called "gillesombud" for a period of about 40 years. The other type was objective estimations based on random samples taken only for the last few years. It was found that the objective estimation method gave considerably higher results than the subjective one, and it was desirable to investigate if the variances behaved in the same way. As a part of this investigation the variation between years and between counties ("län") was studied for some cereals, the yields of which were estimated by the objective method. The observations consisted of one mean yield per "hektar" for every county and year. For the true mean yield in county I and year J we obtain the model:

$$\mu(IJ) = \mu + \alpha(I) + \beta(J) + \gamma(IJ); \quad \begin{matrix} I = 1, 2 \ldots A \\ J = 1, 2 \ldots B \end{matrix}$$

As only a sampling investigation of the mean yield per county and year is made, large technical errors have to be included if we try to observe combination IJ:

$$y(IJ) = \mu + \alpha(I) + \beta(J) + \gamma(IJ) + e(IJ).$$

If the method of estimation is unbiased the random variables $e(IJ)$ will have mean values $= 0$, but the variances $\sigma^2_{e_{IJ}}$ will vary as the number of farms per county, the cereal acreage per farm and even the sizes of the samples vary.

If only a randomly chosen counties and b randomly "chosen" years are observed the statistical model will be:

$$y_{ij} = \mu + \sum_I a_i(I)\,\alpha(I) + \sum_J b_j(J) \cdot \beta(J) + \sum_{IJ} a_i(I)\,b_j(J)\left\{\gamma(IJ) + e(IJ)\right\}$$

The following expected mean squares are obtained:

$$\mathrm{E(MS}_A) = \sigma^2_e + \left(1 - \frac{b}{B}\right)\sigma^2_{AB} + b\sigma^2_A$$

$$\mathrm{E(MS}_B) = \sigma^2_e + \left(1 - \frac{a}{A}\right)\sigma^2_{AB} + a\sigma^2_B$$

$$\mathrm{E(MS}_{AB}) = \sigma^2_e + \sigma^2_{AB}$$

where

$$\sigma^2_e = \frac{1}{AB}\sum_{IJ}\sigma^2_{e_{IJ}}$$

The b years are considered as sampled from an infinitely large population of years (weather conditions) while $a = A$. It is true that not all counties were included, but those rejected were counties in the northern part of Sweden with little or no cultivation of the studied cereals. Thus the factor $\mathcal{A}$ has to be considered fixed. Therefore:

$$\begin{aligned} \mathrm{E}(\mathrm{MS}_A) &= \sigma_e^2 + \sigma_{AB}^2 + b\sigma_A^2 \\ \mathrm{E}(\mathrm{MS}_B) &= \sigma_e^2 + \qquad a\sigma_B^2 \\ \mathrm{E}(\mathrm{MS}_{AB}) &= \sigma_e^2 + \sigma_{AB}^2 \end{aligned}$$

From this we obtain

$$\hat{\sigma}_A^2 = \frac{\mathrm{MS}_A - \mathrm{MS}_{AB}}{b}$$

The variance for the technical errors, which here include sampling and measurement errors, was computed separately for each county and year. Thus it was possible to average these over the a counties and the b years and obtain an estimate of σ_B^2. In a similar manner variance estimates for the subjective method were obtained.

Example 5. Estimation of variances between individuals, and over time within individuals by means of observations on paired organs.[1]

A certain characteristic, measured as a resistance in the eyes, was found to vary both between and within individuals. The interesting feature of this example is that the measurements are made on the eyes, on paired organs. We thereby obtain another population structure than when we measure some characteristic of the blood, the temperature of the body etc.

The treatment of this example can of course be applied to any paired organs. As the human body as well as animal bodies are symmetric in many respects, paired organs are not unusual. Lungs and kidneys etc. occur in pairs.

Let us first discuss the population model in detail. Every individual is assumed to have a certain "quantity" of the characteristic; the "value" for the Ith individual denoted by $p_I(T)$ as a function of time. The characteristic gives rise to a "variable" y, depending on p, that is measurable in the eyes. The values for the two eyes of an individual are at least partly determined at random, and therefore it may be motivated to reckon with a potentially infinite number of eyes, the pair in question constituting a

[1] This problem was posed by Professor Ernst Bárány (2) with whom I have had many fruitful discussions.

random sample. Let us introduce $\mu_{IJ}(T)$ for the true function of time of the Ith individual's Jth eye. It is now useful to redefine the value of the characteristic for individual I as $\mu_{I\cdot}(T)$, which is not necessarily the same as the unmeasurable $p_I(T)$.

Of primary interest is to estimate the variance between the time averages $\mu_{I\cdot}(\cdot)$ of the individuals and the variances for the time-fluctuations, that is for $\mu_{I\cdot}(T)$.

Observations can only be made at discrete time-points, moments, and therefore we introduce the following notation for the true value of variable y for individual no. I, eye no. J and moment no. K:

$$\mu(IJK); \quad \begin{matrix} I = 1, 2 \ldots \infty \\ J = 1, 2 \ldots \infty \\ K = 1, 2 \ldots \infty \end{matrix}$$

These true values can be partitioned into components as follows

$$\mu(IJK) = \mu + \alpha(I) + \beta(IJ) + \gamma_1(IK) + \gamma_2(IJK) \tag{II.5}$$

where $\mu = \mu(\cdot\cdot\cdot)$ is the time average of the population;

$\alpha(I) = \mu(I\cdot\cdot) - \mu$ $(\cdot\cdot\cdot)$ is the deviation between the time averages of the Ith individual and of the population;

$\beta(IJ) = \mu(IJ\cdot) - \mu(I\cdot\cdot)$ is the deviation between the time averages of the IJth eye and of the Ith individual;

$\gamma_1(IK) = \mu(I\cdot K) - \mu(I\cdot\cdot)$ is the deviation between the averages over the population of eyes for individual I's value at moment K and its time average;

$\gamma_2(IJK) = \{\mu(IJK) - \mu(IJ\cdot)\} - \{\mu(I\cdot K) - \mu(I\cdot\cdot)\}$ is the deviation between the deviation of eye no. IJ from its time average at moment K and $\gamma_1(IK)$.

An observation of y for combination IJK would probably result in a measurement error. We write

$$y(IJK) = \mu(IJK) + e(IJK).$$

Let us consider an experiment performed on monkeys. It was not thought safe to make more than one determination per eye. Therefore it was not possible to separate the variance components corresponding to $\beta(IJ)$, $\gamma_2(IJK)$ and $e(IJK)$. If they are added we have the simpler model:

$$y(IK) = \mu + \alpha(I) + \gamma_1(IK) + e(IK)$$

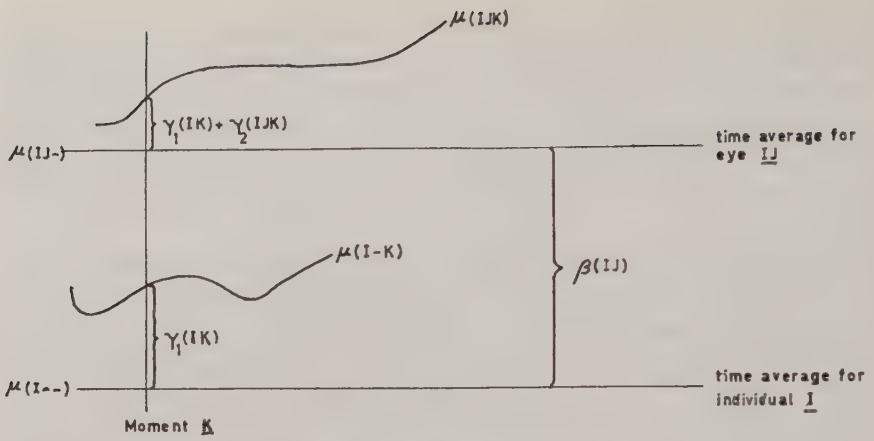


Fig. II: 1. Illustrations of the different components in formula (II.5).

where $e(IK)$ reflects the random error we commit when we try to observe the parameter $\mu_{I\cdot}(T)$ via randomly chosen eyes.

Two types of data were at hand. In the first both eyes of a monkeys had been sampled. Eyes from the same monkey were measured at the same moment, while different animals were (assumed to be) measured at randomly chosen moments.[1]

An analysis between and within for these data will have the following expected mean squares:

$$E\{\mathrm{MS}_A^{(1)}\} = \sigma_e^2 + 2(\sigma_{\gamma_1}^2 + \sigma_A^2)$$

$$E\{\mathrm{MS}_W^{(1)}\} = \sigma_e^2$$

In the second type of data a' monkeys were sampled. The $2 \cdot a'$ eyes were measured at $2 \cdot a'$ randomly chosen moments. In that case the analysis between and within will result in the expected mean squares:

$$E\{\mathrm{MS}_A^{(2)}\} = \sigma_e^2 + \sigma_{\gamma_1}^2 + 2\sigma_A^2$$

$$E\{\mathrm{MS}_W^{(2)}\} = \sigma_e^2 + \sigma_{\gamma_1}^2$$

The estimates of the variance components are thus:

$$\delta_A^2 = \frac{\mathrm{MS}_A^{(2)} - \mathrm{MS}_W^{(2)}}{2}$$

[1] Whether the moments can be considered as independent or not is an important problem not treated here.

$$\hat{\sigma}^2_{\gamma_1} = \frac{\{MS^{(1)}_A - MS^{(1)}_W\} - \{MS^{(2)}_A - MS^{(2)}_W\}}{2}$$

$$\hat{\sigma}^2_e = MS^{(1)}_W$$

The variance of time-fluctuations, $\sigma^2_{\gamma_1}$, is pooled over all individuals in the population. In this investigation it is not possible to estimate the variances for each separate individual.

By considering an infinite number of potential eyes per individual we have developed

(i) a somewhat more satisfying definition of the characteristic $\mu_{I\cdot}(T)$
(ii) a simpler method of estimating the variance for the time-fluctuations.

The latter can be seen in the more general case where we observe the two eyes at c random moments for each individual. "Moments" and "eyes" are then crossed within individuals, and the expected mean squares are easily obtained from Cornfield–Tukey's rules:

$$E(MS_A) = \sigma^2_e + \left(1-\frac{b}{B}\right)\left(1-\frac{c}{C}\right)\sigma^2_{ABC} + c\left(1-\frac{b}{B}\right)\sigma^2_{AB} + b\left(1-\frac{c}{C}\right)\sigma^2_{AC} + bc\sigma^2_A$$

$$E(MS_{AB}) = \sigma^2_e + \left(1-\frac{c}{C}\right)\sigma^2_{ABC} + c\sigma^2_{AB}$$

$$E(MS_{AC}) = \sigma^2_e + \left(1-\frac{b}{B}\right)\sigma^2_{ABC} + b\sigma^2_{AC}$$

$$E(MS_{ABC}) = \sigma^2_e + \sigma^2_{ABC}$$

where factor $\mathcal{A}$ is "individuals", factor $\mathcal{B}$ "eyes" and factor $\mathcal{C}$ "moments". The factors $\mathcal{B}$ and $\mathcal{C}$ are crossed and their combination nested within $\mathcal{A}$. Thus σ^2_{AC} is the variance for the time-fluctuations also denoted by $\sigma^2_{\gamma_1}$. When B is infinite σ^2_{AC} is estimated by $(MS_{AC} - MS_{ABC})/2$, when $B=2$ the variance cannot be estimated unless we make more than one observation per eye per moment so that σ^2_e can be estimated.

2.3. Unbalanced samples

2.3.1. *Introduction*

As has been said before, the numbers of observations in the subclasses are different in an unbalanced sample. Also the number of levels for factor $\mathcal{B}$, nested within factor $\mathcal{A}$, may be different for different levels of $\mathcal{A}$. However, we do not allow the numbers of observations in the subclasses or the

number of levels to be zero; the design (or survey) is complete. Neither shall we consider such small deviations from a balanced sample so that ordinary missing-plot-techniques can be applied.

Unbalanced situations occur frequently in practice, and they are known to make analysis and interpretation difficult. Why do they occur? Why are they not avoided? Let us consider some examples in answer.

Example 1. Estimation of average crop yields.

In some of the investigations discussed in principle in example 1 on page 35, results were obtained for a farms, b_i fields for farm no. i and c_{ij} plots for field no. ij. These results were obtained because sampling was heavier from large farms than from small farms, and in proportion to the sizes of the fields. Thus a smaller variance could be obtained in estimating the average yield. If the farms are randomly sampled from a population of farms, the numbers of fields to be investigated are obviously random variables. For instance, the number of fields to investigate for farm i in the sample depends on which farm in the population has been chosen as the ith farm for the sample. We say that the numbers are determined by the population, not by the sample.

Example 6. Studies of the fluorescence of the eye lens in alloxan diabetic rats.[1]

It is well known that many animal experiments become unbalanced.

In one of the experiments on alloxan diabetic rats, about 75% of the rats were given subcutaneous injections of alloxan monohydrate while the rest were kept as a control group. The reason why these groups could not be kept alike is that many of the injected animals die. I quote from (12): "One week after the diabetogenic dose 38 per cent of the rats died and in all 50 per cent of the injected animals succumbed". Other contributing reasons for the unbalanced situation were:

1) Animals which had not got glucosuria 14 days after the injection were divided into two groups, one which was kept as a control group and one which was reinjected. Some of the reinjected did not show glucosuria and had to be discarded.
2) Both the control and experimental group were divided into two groups. If there were an odd number of rats, one of the treatment groups received one rat more than the other.
3) Some of the animals died because of overdoses when being treated.

[1] Experiments performed by H. Hammar M.D. (12).

4) After the treatment periods the eye lenses were investigated by a complicated method prone to failure. This meant that some of the observations could simply not be made.
5) In some of the experiments the diabetic rats were classified in two groups, with and without cataract, during the course of the investigation.

In such experiments the numbers of observations must be said to be random. That some of the animals are killed by the injections or by overdoses is not as serious here as it could be in other cases when the treatments influence the number of observations (ex. 9). We want to compare healthy with sick (not dead) rats.

Example 7. Studies of the number of cotyledon for pines.[1]

Forty grains of seed were sampled from each of 12 families (four mothers crossed with three fathers). The seed was sown and afterwards the number of cotyledon developed per plant was counted. Data was obtained from between 25 and 39 plants per family, although all families started out with 40 grains of seed. Some of the grains had disappeared when measuring their weight, some were not fed, others did not grow, and some shoots died.

Here the numbers of survivors are determined partly at random and partly by family.

Example 8. Investigating some characteristics of grasshoppers.[2]

In three districts six localities were searched for the insect. The number of grasshoppers to be measured per site was to be limited to those which could be found within a reasonable time. Upper and lower limits were put to 30 and 15 respectively. The sampled number per site will be at least partly proportional to the population number, that is, the total number of insects per site.

Example 9. Comparing the luminous flux for different batches of lightbulbs.[3]

The flux was measured after a burning time of 1 hour and 750 hours. Of course the number of bulbs still burning after 750 hours varied from batch to batch. It could be misleading to compare between the surviving bulbs only, as we can suspect the already burnt-out bulbs of having had

[1] Part of a work by Mr. L. G. Lantz (19).

[2] An application discussed with D. Gärdefors F.l. (11).

[3] Performed at "Statens Institut för konsumentfrågor".

the lowest luminous flux. This is a case where the treatment apparently influences the number of observations. Therefore efforts were made to compute probable values for the burnt-out bulbs. To give these bulbs a flux of 0 would engender underestimations.

Schematically unbalanced situations could be divided into three different types. It will simplify notation but introduce no restriction if we consider a crossed two-factor design with population levels IJ, sample levels ij and cell-frequencies r_{ij}. The three types are then three different assumptions about the r_{ij}:

1) r_{ij} fixed (determined by the sample combination ij),
2) r_{ij} sampled (determined by the population combination IJ, that happened to be sample combination ij),
3) r_{ij} random variables, not completely determined by the population combination IJ.

The same assumptions could of course be made for the levels of a nested factor.

Case 1) is often unrealistic in practice, but it could be relevant, for example, in agricultural experimentation. Wilk–Kempthorne (36) have mentioned that if one factor is "different sites", it may sometimes be desirable, because of economic reasons, to estimate the variance among experimental plots from the first chosen site only, and thus have $r_{i1}=r>1$ but $r_{ij}=1$ for $j>1$. If the accessible number of plots is not a multiple of the number of treatments and one still wants to use all the plots we have another application of case 1). Still another example is given in part IV, page 224.

In sampling investigations the r_{ij}'s are frequently determined by the population combination IJ (example 1).

Case 3) is often relevant when animals die accidentally, analyses fail, observations cannot be made etc.

In cases 2) and 3) we say that we have a situation randomly unbalanced.

2.3.2. *Some notes about the literature on unbalanced situations*

For the classical fixed case the optimum choice of mean squares can be determined by the method of least squares (earlier sometimes called Yate's method of fitting constants). A description can be found in Scheffé (24). Other methods, with different weighting systems, have been proposed. These methods are really only special cases of the method of least squares. Se for example Kramer (18).

However, this exact method is computationally very tedious when one wants to use interaction sums of squares. Therefore it is now often recommended in text books to use another, more rapid, method of analysis. As in all methods the first step is to compute the sum of squares for "within treatment combinations". Then the cell means are analyzed as if they were the only observations. The resulting mean squares have to be multiplied by the harmonic mean of the cell-frequencies to be comparable with the within-mean-square.

If we want to estimate variance components in models other than the fixed model, the method of defining the components and computing the mean squares has to be decided upon. The expected mean squares can then be derived. Several methods have been discussed by Henderson (13). Among other methods there is one where he computes the sums of squares as in a balanced case but with weighted means, ignoring the fact that these SS do not add up to the total. To illustrate the method, let us assume a crossed two-way design with observations

$$y_{ijk};\quad \begin{matrix} i = 1, 2 \ldots a \\ j = 1, 2 \ldots b \\ k = 1, 2 \ldots r_{ij} \end{matrix}$$

The means are

$$y_{ij.} = \frac{1}{r_{ij}} \sum_{k=1}^{r_{ij}} y_{ijk};\quad y_{i..} = \frac{\sum_{j,k} y_{ijk}}{\sum_j r_{ij}};\quad y_{.j.} = \frac{\sum_{i,k} y_{ijk}}{\sum_i r_{ij}};\quad y_{...} = \frac{\sum_{ijk} y_{ijk}}{\sum_{ij} r_{ij}}$$

and the sums of squares

$$SS_A = \sum_{ij} r_{ij}(y_{i..} - y_{...})^2$$

$$SS_B = \sum_{ij} r_{ij}(y_{.j.} - y_{...})^2$$

$$SS_{AB} = \sum_{ij} r_{ij}(y_{ij.} - y_{i..} - y_{.j.} + y_{...})^2$$

When the cell-frequencies are proportional, that is when

$$r_{ij} = \frac{(\sum_i r_{ij}) \cdot (\sum_j r_{ij})}{\sum_{ij} r_{ij}}$$

this analysis is orthogonal in that the sums of squares are additive:

$$SS_A + SS_B + SS_{AB} + SS_{\text{Within}} = SS_{\text{Total}}$$

For the above-mentioned analysis of the cell means the means are defined as follows

$$y_{ij.} = \frac{1}{r_{ij}} \sum_k y_{ijk}; \quad y_{i..} = \frac{1}{b} \sum_j y_{ij.}; \quad y_{.j.} = \frac{1}{a} \sum_i y_{ij.}; \quad y_{...} = \frac{1}{ab} \sum_{ij} y_{ij.}$$

and the sums of squares

$$SS_A = \frac{ab}{\sum_{ij}(1/r_{ij})} \sum_{ij} (y_{i..} - y_{...})^2$$

$$SS_B = \frac{ab}{\sum_{ij}(1/r_{ij})} \sum_{ij} (y_{.j.} - y_{...})^2$$

$$SS_{AB} = \frac{ab}{\sum_{ij}(1/r_{ij})} \sum_{ij} (y_{ij.} - y_{i..} - y_{.j.} + y_{...})^2$$

These two analyses will in the following be referred to as the weighted analysis and unweighted analysis.

The method of least squares assuming fixed effects is also used at times. In random cases, which in this differ from the fixed case, it is not possible to tell which method is the better. In such a situation it would be preferable to have mean squares that are easy to compute and whose expectations are readily calculated.

Abt (1) has given expected mean squares for the weighted analysis with proportional frequencies when there are two, three or four crossed factors. In his approach the plot-population is always infinite; he does therefore not need to distinguish between the sampling and experimental cases. The factors are either fixed or random, that is, always $Q = q$ or Q infinite.

Wilk–Kempthorne (36, 37) have computed E(MS)'s for the weighted and unweighted analysis using proportional[1] and general frequencies respectively. They consider a completely randomized design with three crossed factors. The factor levels are considered sampled from finite populations, but the cell-frequencies are fixed (case 1) on page 64).

Randomly unbalanced situations, that is, when the cell-frequencies are random variables (cases 2) and 3)), I have seen discussed only in literature on sampling. Sukhatme (31) considers the case of two-stage sampling, with the numbers of observations in the second stage completely determined by the population. He defines a sort of mean squares between and within primary sampling units, in a way that will be used in the next paragraph. However, Sukhatme does not call this method analysis of variance. Some pages later he uses the weighted analysis of variance to estimate the variance of his estimate under the restrictions that the cell-

[1] Some differences between their and Abt's results will be commented upon on page 77.

frequencies are fixed and the population balanced. In a comment to the latter method he points out "that the system of drawing m_i second-stage units at the ith draw, irrespective of which first-stage unit is included in the sample at the ith draw, is not a rational system which is likely to be used in practice". When the cell-frequencies m_i are completely random, he says, and when the population is only slightly unbalanced and the number of first-stage units in the population is large then the anova-method of estimation could be used as a sufficiently good approximation.

However, in this work Sukhatme's first mentioned method is called analysis of variance, which is said to be a method for splitting a total variance into components, that do not necessarily sum up to the total.

2.3.3. *Modified analysis with a randomly unbalanced sample*

The method easiest to handle in unbalanced situations is that with unweighted means and it will be used in the following. In general it is not possible to say whether the weighted or unweighted analysis is the best. What is considered important here is that the unweighted analysis suitably modified can be shown to give unbiased estimates (under certain restrictions) of the variance components. Let us begin by considering the simplest assumption of the cell-frequencies, namely case 2).

a) *Illustrations from sampling.*—In a sampling case with two factors, $\mathcal{A}$ nesting $\mathcal{B}$, we could use the following notations in the population:

Levels of factor $\mathcal{A}$	$\overbrace{1}$	$\overbrace{2}$	...	$\overbrace{I}$	...	$\overbrace{A}$
Levels of factor $\mathcal{B}$	$\overbrace{1,2 \dots J \dots B_1}$	$\overbrace{1,2 \dots J \dots B_2}$		$\overbrace{1,2 \dots J \dots B_I}$		$\overbrace{1,2 \dots J \dots B_A}$
Number of levels of $\mathcal{B}$ to be chosen	b_1	b_2		b_I		b_A

Model:

$$\mu(IJ) = \mu + \alpha(I) + \beta(IJ)$$

$$\sum_J \beta(IJ) = \sum_I \alpha(I) = 0$$

If a levels of factor A and the corresponding numbers of levels of factor $\mathcal{B}$ are chosen at random the statistical model (disregarding technical errors) can be written

$$y_{ij} = \mu + \sum_I a_i(I)\,\alpha(I) + \sum_{IJ} a_i(I)\,b_{Ij}(IJ)\cdot\beta(IJ); \quad \begin{matrix} i = 1, 2 \dots a \\ j = 1, 2 \dots b_i \end{matrix}$$

The analysis of unweighted means will use

$$y_{i.} = \frac{1}{b_i}\sum_j y_{ij} = \mu + \sum_I a_i(I)\,\alpha(I) + \sum_{IjJ} a_i(I)\cdot\frac{1}{b_I}\,b_{Ij}(IJ)\cdot\beta(IJ)$$

$$y_{..} = \frac{1}{a}\sum_i y_{i.} = \mu + \frac{1}{a}\sum_{iI} a_i(I)\,\alpha(I) + \frac{1}{a}\sum_{iIjJ} a_i(I)\,\frac{1}{b_I}\cdot b_{Ij}(IJ)\cdot\beta(IJ)$$

$$\mathrm{MS}_A = \frac{1}{a-1}\sum_i (y_{i.} - y_{..})^2$$

$$\mathrm{MS}''_{AB} = \frac{\sum_i (1/b_i)}{a}\cdot\frac{1}{\sum_i (b_i - 1)}\sum_{ij} (y_{ij} - y_{i.})^2$$

Instead of multiplying MS_A with the harmonic mean of the cell-frequencies as on page 66 we have here divided the within-mean-square.

The next step is to derive the expected mean squares. $\underset{2}{E}$ will here mean expectation over the second stage, that is given the chosen a levels of factor $\mathcal{A}$; $\underset{1,2}{E}$ will mean expectation over both stages.

$$E\{(a-1)\,\mathrm{MS}_A\} = E\sum_i\left[\sum_I a_i(I)\,\alpha(I) + \sum_{IjJ} a_i(I)\,\frac{1}{b_I}\cdot b_{Ij}(IJ)\cdot\beta(IJ) - \frac{1}{a}\sum_{iI} a_i(I)\,\alpha(I)\right.$$
$$\left. - \frac{1}{a}\sum_{iIjJ} a_i(I)\,\frac{1}{b_I}\cdot b_{Ij}(IJ)\cdot\beta(IJ)\right]^2 = E\sum_i\left[\sum_I a_i(I)\,\alpha(I) - \frac{1}{a}\sum_{iI} a_i(I)\,\alpha(I)\right]^2$$
$$+ E\sum_i\left[\sum_{IjJ} a_i(I)\,\frac{1}{b_I}\cdot b_{Ij}(IJ)\cdot\beta(IJ) - \frac{1}{a}\sum_{iIjJ} a_i(I)\cdot\frac{1}{b_I}\cdot b_{Ij}(IJ)\cdot\beta(IJ)\right]^2$$
$$+ \text{a vanishing crossproduct} = E(A_1) + E(A_2)$$

$$E\{A_1\} = E\left\{\sum_i\sum_I a_i(I)\alpha^2(I)\right\} - \frac{1}{a}\left[E\left\{\sum_{iI} a_i(I)\alpha^2(I)\right\} + E\left\{\sum_{i\neq i'\,I\neq I'} a_i(I)a_{i'}(I')\alpha(I)\alpha(I')\right\}\right]$$
$$= a\cdot\frac{1}{A}\sum_I \alpha^2(I) - \frac{1}{a}\left\{a\cdot\frac{1}{A}\sum_I\alpha^2(I) - a(a-1)\,\frac{1}{A(A-1)}\sum_I\alpha^2(I)\right\}$$
$$= (a-1)\,\frac{1}{A-1}\sum_I\alpha^2(I) = (a-1)\,\sigma_A^2$$

$$E\{A_2\} =$$
$$E\sum_i\left[\sum_{IjJ} a_i(I)\,\frac{1}{b_I^2}\cdot b_{Ij}(IJ)\,\beta^2(IJ) + \sum_{Ij\neq j'J\neq J'} a_i(I)\,\frac{1}{b_I^2}\cdot b_{Ij}(IJ)b_{Ij'}(IJ')\,\beta(IJ)\beta(IJ')\right]$$
$$-\frac{1}{a}\left[E\left\{\sum_{iIjJ} a_i(I)\frac{1}{b_I^2}b_{Ij}(IJ)\,\beta^2(IJ)\right\} + E\left\{\sum_{\substack{iIj\neq j'\\ J\neq J'}} a_i(I)\frac{1}{b_I^2}b_{Ij}(IJ)b_{Ij'}(IJ')\beta(IJ)\beta(IJ')\right\}\right]$$
$$-\frac{1}{a}E\left\{\sum_{\substack{i\neq i'\,I\neq I'\\ jj'JJ'}} a_i(I)\,a_{i'}(I')\,\frac{1}{b_I}\cdot\frac{1}{b_{I'}}\cdot b_{Ij}(IJ)\,b_{Ij'}(I'J')\,\beta(IJ)\,\beta(I'J')\right\}$$

The last term vanishes because $\sum_I \beta(IJ) = 0$. Therefore one obtains

$$E(A_2) = E(A_{21}) + E(A_{22}) - \frac{1}{a}E(A_{23}) - \frac{1}{a}E(A_{24})$$

$$E(A_{21}) = \underset{1}{E}\left[\sum_{iI} a_i(I)\cdot\frac{1}{b_I^2}\underset{2}{E}\left\{\sum_{jJ} b_{Ij}(IJ)\cdot\beta^2(IJ)\right\}\right] = \underset{1}{E}\left\{\sum_{iI}\left[a_i(I)\cdot\frac{1}{b_I^2}\cdot b_I\cdot\frac{1}{B_I}\sum_J\beta^2(IJ)\right]\right\}$$

$$= a\cdot\frac{1}{A}\sum_I\frac{1}{b_I}\cdot\frac{1}{B_I}\sum_J\beta^2(IJ)$$

$$E(A_{22}) = \underset{1}{E}\sum_{iI}\left[a_i(I)\frac{1}{b_I^2}\underset{2}{E}\left\{\sum_{J\neq J'J\neq J'} b_{Ij}(IJ)\,b_{Ij'}(IJ')\,\beta(IJ)\cdot\beta(IJ')\right\}\right]$$

$$= -a\frac{1}{A}\sum_I(b_I-1)\frac{1}{b_I}\frac{1}{B_I}\cdot\frac{1}{B_I-1}\sum_J\beta^2(IJ)$$

$$E(A_{23}) = a\frac{1}{A}\sum_I\frac{1}{b_I}\frac{1}{B_I}\cdot\sum_J\beta^2(IJ)$$

$$E(A_{24}) = -a\cdot\frac{1}{A}\sum_I(b_I-1)\frac{1}{b_I}\frac{1}{B_I}\cdot\frac{1}{B_I-1}\sum_J\beta^2(IJ)$$

$$E(A_2) = (a-1)\frac{1}{A}\sum_I\frac{1}{b_I}\frac{1}{B_I}\sum_J\beta^2(IJ) - (a-1)\frac{1}{A}\sum_I\frac{1}{b_I}\frac{b_I-1}{B_I-1}\frac{1}{B_I}\sum_J\beta^2(IJ)$$

$$= (a-1)\cdot\frac{1}{A}\sum_I\left(\frac{1}{b_I}-\frac{1}{B_I}\right)\frac{1}{B_I-1}\sum_J\beta^2(IJ) = (a-1)\frac{1}{A}\sum_I\left(\frac{1}{b_I}-\frac{1}{B_I}\right)\sigma^2_{AB_I}$$

$$\because\quad \mathrm{E(MS}_A) = \frac{1}{A}\sum_I\left(\frac{1}{b_I}-\frac{1}{B_I}\right)\sigma^2_{AB_I} + \sigma^2_A$$

Thus one met with no difficulty in deriving the expected mean square between the levels of factor $\mathcal{A}$. The mean square within, however, has the product between two random variables in the numerator and another random variable in the denominator. Calculating expectation for this would be quite difficult, and besides, it would not result in $(1/A)\Sigma_I(1/b_I - 1/B_I)\sigma^2_{AB_I}$ which is desirable if one wants an unbiased estimate of σ^2_A.

We therefore modify the mean square in the following way

$$\mathrm{MS}_{AB} = \frac{1}{a}\sum_{ij}\left(\frac{1}{b_i}-\frac{1}{B_i}\right)\frac{1}{b_i-1}(y_{ij}-y_{i.})^2$$

$$E\{a\cdot \mathrm{MS}_{AB}\} = E\sum_{ij}\left[\left\{\sum_I a_i(I)\left(\frac{1}{b_I}-\frac{1}{B_I}\right)\cdot\frac{1}{b_I-1}\right\}\left\{\sum_{IJ} a_i(I)\,b_{Ij}(IJ)\,\beta(IJ)\right.\right.$$

$$\left.\left.-\sum_{IJ} a_i(I)\,\frac{1}{b_I}\cdot b_{Ij}(IJ)\cdot\beta(IJ)\right\}^2\right] = E\sum_{ij}\left[\sum_I a_i(I)\left(\frac{1}{b_I}-\frac{1}{B_I}\right)\cdot\frac{1}{b_I-1}\right.$$

$$\left.\times\left\{\sum_J b_{Ij}(IJ)\,\beta(IJ)-\frac{1}{b_I}\sum_{jJ} b_{Ij}(IJ)\,\beta(IJ)\right\}^2\right] = E_1\sum_{iI}\left[a_i(I)\left(\frac{1}{b_I}-\frac{1}{B_I}\right)\right.$$

$$\left.\times\frac{1}{b_I-1}E_2\sum_j\left\{\sum_J b_{Ij}(IJ)\,\beta(IJ)-\frac{1}{b_I}\sum_{jJ} b_{Ij}(IJ)\,\beta(IJ)\right\}^2\right]$$

$$= E_1\sum_{iI}\left[a_i(I)\left(\frac{1}{b_I}-\frac{1}{B_I}\right)\cdot\frac{1}{b_I-1}\left\{E_2\sum_{jJ} b_{Ij}(IJ)\,\beta^2(IJ)\right.\right.$$

$$\left.\left.-\frac{1}{b_I}E_2\left[\sum_{jJ} b_{Ij}(IJ)\,\beta^2(IJ)+\sum_{j\neq j'J\neq J'} b_{Ij}(IJ)\,b_{Ij'}(IJ')\,\beta(IJ)\,\beta(IJ')\right]\right\}\right]$$

$$= a\frac{1}{A}\sum_I\left[\left(\frac{1}{b_I}-\frac{1}{B_I}\right)\frac{1}{b_I-1}\left\{b_I-1+(b_I-1)\frac{1}{B_I-1}\right\}\frac{1}{B_I}\sum_J\beta^2(IJ)\right]$$

$$= a\frac{1}{A}\sum_I\left(\frac{1}{b_I}-\frac{1}{B_I}\right)\sigma^2_{AB_I}$$

Thus
$$\mathrm{E}(\mathrm{MS}_{AB}) = \frac{1}{A}\sum_I\left(\frac{1}{b_I}-\frac{1}{B_I}\right)\sigma^2_{AB_I}$$

and
$$\hat{\sigma}^2_A = \mathrm{MS}_A - \mathrm{MS}_{AB}$$

An unbiased estimate of $\sigma^2_{AB} = (1/A)\,\Sigma_I\,\sigma^2_{AB_I}$ is obtained by

$\hat{\sigma}^2_{AB} = \mathrm{MS}'_{AB} = \frac{1}{a}\sum_i\frac{1}{b_i-1}(y_{ij}-y_{i\cdot})^2$ as it is easy to show that $\mathrm{E}(\mathrm{MS}'_{AB}) = \sigma^2_{AB}$

When the number of factors is large several modifications of the mean squares could be used. Let us, for example, consider a sampling situation where factor $\mathcal{A}$ nests factor $\mathcal{B}$, which in turn nests factor $\mathcal{C}$. The population notations are

Levels of factor $\mathcal{A}$	$\overbrace{\qquad\qquad}^{1}$	$\overbrace{\qquad\qquad}^{2}$	...	$\overbrace{\qquad\qquad}^{I}$	...	$\overbrace{\qquad\qquad}^{A}$
Levels of factor $\mathcal{B}$	$\overbrace{1,2\ldots J\ldots B_1}$	$\overbrace{1,2\ldots J\ldots B_2}$		$\overbrace{1,2\ldots J\ldots B_I}$		$\overbrace{1,2\ldots J\ldots B_A}$
Levels of factor $\mathcal{C}$	1 1			1		
	2 2			2		
	⋮ ⋮			⋮		
	K K			K		
	⋮ ⋮			⋮		
	C_{11} C_{1J}			C_{IJ}		

The number of levels to be chosen for factor $\mathcal{B}$ is b_I and for factor $\mathcal{C}$ it is c_{IJ}.

The observations are denoted by

$$y_{ijk};\quad \begin{matrix} i = 1, 2 \ldots a \\ j = 1, 2 \ldots b_i \\ k = 1, 2 \ldots c_{ij} \end{matrix}$$

If the sampling is to be done with equal probabilities we define the following mean squares:

$$\mathrm{MS}_A = \frac{1}{a-1} \sum_i (y_{i..} - y_{...})^2$$

$$\mathrm{MS}_{AB} = \frac{1}{a} \sum_{ij} \frac{1}{b_i - 1} (y_{ij.} - y_{i..})^2$$

$$\mathrm{MS}'_{AB} = \frac{1}{a} \sum_{ij} \frac{1}{b_i(b_i - 1)} (y_{ij.} - y_{i..})^2$$

$$\mathrm{MS}_{ABC} = \frac{1}{a} \sum_{ijk} \frac{1}{b_i(c_{ij} - 1)} (y_{ijk} - y_{ij.})^2$$

and obtain

$$\mathrm{E}(\mathrm{MS}_A) = \frac{1}{A} \sum_I \frac{1}{b_I} \frac{1}{B_I} \sum_J \frac{1}{c_{IJ}} [\sigma^2_{e_{IJ}} + \sigma^2_{ABC_{IJ}}]$$
$$+ \frac{1}{A} \sum_I \left(\frac{1}{b_I} - \frac{1}{B_I} \right) \left[\sigma^2_{AB_I} - \frac{1}{B_I} \sum_J \frac{1}{C_{IJ}} \sigma^2_{ABC_{IJ}} \right] + \left[\sigma^2_A - \frac{1}{A} \sum_I \frac{1}{B_I^2} \sum_J \frac{1}{C_{IJ}} \sigma^2_{ABC_{IJ}} \right]$$

$$\mathrm{E}(\mathrm{MS}_{AB}) = \frac{1}{A} \sum_I \frac{1}{B_I} \sum_J \frac{1}{c_{IJ}} [\sigma^2_{e_{IJ}} + \sigma^2_{ABC_{IJ}}] + \frac{1}{A} \sum_I \left[\sigma^2_{AB_I} - \frac{1}{B_I} \sum_J \frac{1}{C_{IJ}} \cdot \sigma^2_{ABC_{IJ}} \right]$$

$$\mathrm{E}(\mathrm{MS}'_{AB}) = \frac{1}{A} \sum_I \frac{1}{b_I} \frac{1}{B_I} \sum_J \frac{1}{c_{IJ}} [\sigma^2_{e_{IJ}} + \sigma^2_{ABC_{IJ}}] + \frac{1}{A} \sum_I \frac{1}{b_I} \left[\sigma^2_{AB_I} - \frac{1}{B_I} \sum_J \frac{1}{C_{IJ}} \sigma^2_{ABC_{IJ}} \right]$$

$$\mathrm{E}(\mathrm{MS}_{ABC}) = \sigma^2_e + \sigma^2_{ABC}$$

If $1/B_I$ can be neglected compared to $1/b_I$, $\hat{\sigma}^2_A = \mathrm{MS}_A - \mathrm{MS}'_{AB}$ is an unbiased estimate of σ^2_A.

If it is easy to determine B_i for every chosen level of factor $\mathcal{A}$, we obtain the following estimation possibilities:

1) $$\mathrm{MS}'_{ABC} = \frac{1}{a} \sum_{ijk} \frac{1}{b_i B_i c_{ij}(c_{ij} - 1)} (y_{ijk} - y_{ij.})^2$$

$$E(MS'_{ABC}) = \frac{1}{A}\sum_I \frac{1}{B_I}\sum_J \frac{1}{c_{IJ}}[\sigma^2_{e_{IJ}} + \sigma^2_{ABC_{IJ}}]$$

which if $1/C_{IJ} \ll 1/c_{IJ}$ gives

$$\hat{\sigma}^2_{AB} = MS_{AB} - MS'_{ABC}$$

2)
$$MS''_{AB} = \frac{1}{a}\sum_{ij} \frac{1}{b_i - 1}\left(\frac{1}{b_i} - \frac{1}{B_i}\right)(y_{ij.} - y_{i..})^2$$

$$MS''_{ABC} = \frac{1}{a}\sum_{ijk} \frac{1}{b_i B_i}\left(\frac{1}{b_i} - \frac{1}{B_i}\right)\frac{1}{c_{ij}(c_{ij} - 1)}(y_{ijk} - y_{ij.})^2$$

$$MS'''_{ABC} = \frac{1}{a}\sum_{ijk} \frac{1}{b_i^2 B_i} \cdot \frac{1}{c_{ij}(c_{ij} - 1)}(y_{ijk} - y_{ij.})^2$$

the expected values of which when $1/C_{IJ} \ll 1/c_{IJ}$ are

$$E(MS''_{AB}) = \frac{1}{A}\sum_I \frac{1}{B_I}\left(\frac{1}{b_I} - \frac{1}{B_I}\right)\sum_J \frac{1}{c_{IJ}}[\sigma^2_{e_{IJ}} + \sigma^2_{ABC_{IJ}}] + \frac{1}{A}\sum_I \left(\frac{1}{b_I} - \frac{1}{B_I}\right)\sigma^2_{AB_I}$$

$$E(MS''_{ABC}) = \frac{1}{A}\sum_I \frac{1}{B_I}\left(\frac{1}{b_I} - \frac{1}{B_I}\right)\sum_J \frac{1}{c_{IJ}}[\sigma^2_{e_{IJ}} + \sigma^2_{ABC_{IJ}}]$$

$$E(MS'''_{ABC}) = \frac{1}{A}\sum_I \frac{1}{b_I \cdot B_I}\sum_J \frac{1}{c_{IJ}}[\sigma^2_{e_{IJ}} + \sigma^2_{ABC_{IJ}}]$$

and thus

$$\hat{\sigma}^2_A = (MS_A - MS'''_{ABC}) - (MS''_{AB} - MS''_{ABC})$$

without any assumption about $1/B_I$.

In order to obtain an unbiased estimate of σ^2_A it is required *either*, that all B_I are infinite *or* that all C_{IJ} are infinite and B_i easily determined.

The requirements for an unbiased estimate of σ^2_{AB} are that all C_{IJ} are infinite and B_i easily determined.

If one goes through the case with fixed r_{ij} in the same manner, the estimation will be made somewhat easier. It can be shown that an unbiased estimate is obtained for σ^2_A under exactly the same requirements as for sampled r_{ij}, for σ^2_{AB} if *either* all B_I are infinite *or* all C_{IJ} are infinite.

Thus it is possible to define suitable mean squares so that the variance components may be estimated without bias, but the multiplicity of mean squares makes it questionable whether this method of analysis falls under the definition of analysis of variance.

b) *An experimental illustration.*—This type of analysis could as well be used with crossed factors as well as in the experimental case. However, in the experimental case the cell-frequencies are probably not so often determined by the population, at least not in the same way as they have been used above. A possible application would be a randomized block design where the blocks differ in size and we use a different number of plots per block.

Let us consider a randomized block design with one treatment factor. We adopt the model on page 52 with the change that here are P_J plots in block J; $K = 1, 2 \ldots P_J$; which will alter some of the variance components in an obvious manner. The observed values are

$$y_{ijf};\quad \begin{matrix} i = 1, 2 \ldots a \\ j = 1, 2 \ldots b \\ f = 1, 2 \ldots r_{ij} \end{matrix}$$

The mean squares for the unweighted analysis are

$$\mathrm{MS}_A = \frac{1}{a-1} \sum_{ij} (y_{i..} - y_{...})^2$$

$$\mathrm{MS}_B = \frac{1}{b-1} \sum_{ij} (y_{.j.} - y_{...})^2$$

$$\mathrm{MS}_{AB} = \frac{1}{(a-1)(b-1)} \sum_{ij} (y_{ij.} - y_{i..} - y_{.j.} + y_{...})^2$$

$$\mathrm{MS}_{ABP} = \frac{1}{ab} \sum_{ijf} \frac{1}{r_{ij}(r_{ij}-1)} \cdot (y_{ijf} - y_{ij.})^2$$

$$\mathrm{MS}'_{ABP} = \frac{1}{ab} \sum_{ijf} \frac{1}{r_{ij}-1} \cdot (y_{ijf} - y_{ij.})^2$$

In MS_{ABP} the cell variances are weighted in inverse proportion to the cell-frequencies

$$\mathrm{MS}_{ABP} = \frac{1}{ab} \sum_{ij} \frac{1}{r_{ij}} \hat{s}^2_{ij}$$

With the usual within-mean-square they are weighted in direct proportion to the cell variances

$$\mathrm{MS}''_{ABP} = \frac{\sum_{ij} (1/r_{ij})}{ab \sum_{ij} (r_{ij}-1)} \sum_{ij} (r_{ij}-1) \hat{s}^2_{ij}$$

The expected values for these mean squares look rather complicated at first sight. We have even been forced to introduce

$$\mathrm{Cov}_{IJ}(BP;\, ABP) = \frac{1}{P_J - 1} \sum_K \varepsilon(JK) \cdot \eta(IJK)$$

as this term does not vanish. For comparison expected mean squares are given in table II: 6 for both the balanced case and for the unbalanced case, the latter with both fixed and sampled r_{ij}.

From table II: 6 it can be seen that exactly the same components can be estimated and in exactly the same way in any of the three cases. Compared to the ordinary analysis in the balanced case there are small modifications in the estimation formulas for the variance components. For example the component $[\sigma^2_{AB} - A/(A-1)(1/AB)\Sigma_{IJ}(1/P_J)\sigma^2_{ABP_{IJ}}]$ is estimated by $\mathrm{MS}_{AB} - \mathrm{MS}_{ABP}$ instead of by $(\mathrm{MS}_{AB} - \mathrm{MS}_{ABP})/r$ as in the balanced analysis.

It is easy to see from (iii) that if $r_{IJ} = r_J$ for all I[1] $\mathrm{Cov}(BP\,;\, ABP)$ vanishes everywhere and if besides all plots are used σ^2_{BP} vanishes in $\mathrm{E}(\mathrm{MS}_B)$. The latter is not the case when the r_{IJ} within a block are different. At first sight this may seem curious, when all plots in the blocks are used the plot-variance within blocks should not contribute to the variance between blocks, but the reason is that the block means $y_{.j.}$ are unweighted means of the cell means $y_{ij.}$ and thus the plots are weighted differently.

c) *Comments on case 3) regarding assumptions about cell-frequencies.*— At first we assume that every plot randomized to factor combination IJ has equal probability, p_{IJ}, of "survival". Then the actual procedure of randomizing a number of plots, some of which will "die", to combination IJ can be considered equivalent to

1) sampling a number r_{IJ} from a population of numbers and

2) sampling r_{IJ} plots with equal probabilities from the population of plots.

Thus it is possible to calculate conditional expectations step by step with the result that $1/r_{IJ}$ in Table II: 6 (iii) is changed to $E(1/r_{IJ})$, that is, the expectation of the inverted values of all possible numbers in the population of numbers for combination IJ. Note that we only consider complete designs with possible values of $r_{IJ} > 0$.

More often the plots are so different that the probabilities of "survival" differ. In the thought procedure above this corresponds to choosing the r_{IJ} plots for combination IJ with unequal probabilities (in proportion to their "survival" probabilities). However, if the plot populations are infinite, and if we assess different weights to different plots in the definitions of components of variance, the expected mean squares are unchanged.

[1] Which will probably be the most common in practice.

TABLE II: 6. *Expected mean squares in modified unweighted analysis of a randomized block design with one treatment factor A.*

(i) *Balanced case*

$$E(MS_A) = \frac{1}{r}\cdot\sigma^2 + \left(1-\frac{b}{B}\right)\left[\sigma^2_{AB} - \frac{1}{P}\cdot\frac{A}{A-1}\cdot\sigma^2_{ABP}\right] + b\left[\sigma^2_A - \frac{1}{BP}\cdot\frac{A}{A-1}\cdot\sigma^2_{ABP}\right]$$

$$E(MS_B) = \frac{1}{r}\cdot\sigma^2 + \left(1-\frac{a}{A}\right)\left[\sigma^2_{AB} - \frac{1}{P}\cdot\frac{A}{A-1}\cdot\sigma^2_{ABP}\right] + a\left[\sigma^2_B - \frac{1}{P}\cdot\sigma^2_{BP}\right]$$

$$E(MS_{AB}) = \frac{1}{r}\cdot\sigma^2 + \left[\sigma^2_{AB} - \frac{1}{P}\cdot\frac{A}{A-1}\cdot\sigma^2_{ABP}\right]$$

$$E(MS_{ABP}) = \frac{1}{r}\cdot\sigma^2$$

$$E(MS'_{ABP}) = \sigma^2 = \sigma^2_e + \sigma^2_{BP} + \sigma^2_{ABP}$$

(ii) *Unbalanced case, r_{ij} fixed*

$$E(MS_A) = \frac{1}{ab}\sum_{ij}\frac{1}{r_{ij}}\cdot\sigma^2 + \left(1-\frac{b}{B}\right)\left[\sigma^2_{AB} - \frac{A}{A-1}\cdot\frac{1}{AB}\sum_{IJ}\frac{1}{P_J}\cdot\sigma^2_{ABP_{IJ}}\right] + b\left[\sigma^2_A - \frac{1}{B}\cdot\frac{A}{A-1}\cdot\frac{1}{AB}\cdot\sum_{IJ}\frac{1}{P_J}\sigma^2_{ABP_{IJ}}\right]$$

$$E(MS_B) = \frac{1}{ab}\sum_{ij}\frac{1}{r_{ij}}\cdot\sigma^2 + \left(1-\frac{a}{A}\right)\left[\sigma^2_{AB} - \frac{A}{A-1}\cdot\frac{1}{AB}\sum_{IJ}\frac{1}{P_J}\sigma^2_{ABP_{IJ}}\right] + a\left[\sigma^2_B - \frac{1}{B}\cdot\sum_J\frac{1}{P_J}\sigma^2_{BP_J}\right]$$

$$E(MS_{AB}) = \frac{1}{ab}\sum_{ij}\frac{1}{r_{ij}}\cdot\sigma^2 + \left[\sigma^2_{AB} - \frac{A}{A-1}\frac{1}{AB}\sum_{IJ}\frac{1}{P_J}\sigma^2_{ABP_{IJ}}\right]$$

$$E(MS_{ABP}) = \frac{1}{ab}\sum_{ij}\frac{1}{r_{ij}}\cdot\sigma^2$$

$$E(MS'_{ABP}) = \sigma^2$$

(iii) *Unbalanced case, r_{ij} sampled*

$$E(MS_A) = \frac{1}{AB}\sum_{IJ}\frac{1}{r_{IJ}}\cdot\sigma^2_{IJ} + \left(1-\frac{b}{B}\right)\left[\sigma^2_{AB} - \frac{A}{A-1}\cdot\frac{1}{AB}\sum_{IJ}\frac{1}{P_J}\sigma^2_{ABP_{IJ}}\right] + b\left[\sigma^2_A - \frac{1}{B}\cdot\frac{A}{A-1}\cdot\frac{1}{AB}\sum_{IJ}\frac{1}{P_J}\sigma^2_{ABP_{IJ}}\right]$$

Table II: 6 (*continued*).

$$E(MS_B) = \frac{1}{AB}\sum_{IJ}\frac{1}{r_{IJ}}\cdot\sigma^2_{IJ} + \left(1-\frac{a}{A}\right)\left[\sigma^2_{AB} - \frac{A}{A-1}\cdot\frac{1}{AB}\sum_{IJ}\frac{1}{P_J}\sigma^2_{ABP_{IJ}}\right]$$

$$+ a\left[\sigma^2_B - \frac{1}{B}\cdot\sum_J\frac{1}{P_J}\sigma^2_{BP_J}\right]$$

$$E(MS_{AB}) = \frac{1}{AB}\sum_{IJ}\frac{1}{r_{IJ}}\cdot\sigma^2_{IJ} + \left[\sigma^2_{AB} - \frac{A}{A-1}\cdot\frac{1}{AB}\sum_{IJ}\frac{1}{P_J}\sigma^2_{ABP_{IJ}}\right]$$

$$E(MS_{ABP}) = \frac{1}{AB}\sum_{IJ}\frac{1}{r_{IJ}}\cdot\sigma^2_{IJ} = \frac{1}{AB}\sum_{IJ}\frac{1}{r_{IJ}}\left[\sigma^2_{e_{IJ}} + \sigma^2_{BP_J} + 2\mathrm{Cov}_{IJ}(BP;ABP) + \sigma^2_{ABP_{IJ}}\right]$$

$$E(MS'_{ABP}) = \sigma^2 = \sigma^2_e + \sigma^2_{BP} + \sigma^2_{ABP}$$

2.3.4. *Conclusions*

In the foregoing paragraph it has been proposed to use a modified analysis, sometimes used in multi-stage sampling, even in unbalanced experiments.

The method consists of weighting the within-variances in inverse instead of in direct proportion to the cell-frequencies, that is, to use for example

$$MS_{\text{Within}} = \frac{1}{ab}\sum_{ij}\frac{1}{r_{ij}}s^2_{ij}$$

instead of

$$MS_{\text{Within}} = \frac{1}{ab}\frac{\sum_{ij}1/r_{ij}}{\sum_{ij}(r_{ij}-1)}\sum_{ij}(r_{ij}-1)\,s^2_{ij}$$

as in the usual unweighted analysis.

It may seem to be much less efficient to assess less weight to variances from larger groups. This is also true if one wants to estimate a common within-variance, but here we allow the variances to vary from cell to cell, and moreover, we use the modified mean squares to estimate other variance components than the within-variance component.

The reason for the use of this modified analysis was, that it made possible the computation of expected mean squares when the cell-frequencies are random variables. The estimation formulas for the variance components are the same for both fixed and sampled cell-frequencies, and it was argued that this is also the case in an arbitrary random case. Therefore it is recommended to use the modified analysis as soon as the sample becomes unbalanced; then one need not bother about the type of unbalanced situa-

tion that occurs. Of course great caution is necessary if there is any suspicion that the treatments to be compared influence the number of observations.

The expected mean squares in table II: 6 are derived under the assumption of sampling with equal probabilities. The factor levels are also given equal weights. If in sampling levels with equal probabilities and then in computing means or mean squares the levels are weighted according to importance (for example, when the levels are sites of different importance in an agricultural experiment), products and quotients between random variables are reintroduced. In example 1 the levels are given unequal weights, but as the weights are in proportion to the sampling probabilities, they are automatically introduced in the mean squares and thus cause no problem.

For the case of proportional cell-frequencies Abt (1) weights the levels in proportion to marginal sums of the numbers of observations. In a mixed two-way design with the marginal sums $s_1, s_2 \ldots s_a$ for the levels of the fixed factor $\mathcal{A}$ he defines, for example, $\mu(-J) = \Sigma_I s_I \mu(IJ)/\Sigma_I s_I$. This could be motivated if the levels of factor $\mathcal{A}$ were important in proportion to the s_I, but not otherwise. Wilk–Kempthorne (36) have also used the weighted analysis for proportional frequencies but have defined unweighted population means, $\mu(-J) = (1/a)\Sigma_I \mu(IJ)$. As the levels are weighted differently in the analysis, this means that σ^2_{AB} does not vanish in the E(MS) for the random factor in a mixed model, which it does in Abt's case and when using the unweighted analysis. The same problem have been discussed in connection with table II: 6 on page 74. As Abt (1) pointed out, for the case of Wilk–Kempthorne, the contrast in two weighted means for the random factor, $\bar{y}_{.j.} - \bar{y}_{.j'.}$, will also be a biased estimate of $\beta(j) - \beta(j')$, as these effects are unweighted. However, as a rule one is interested in contrasts in the fixed effects $\alpha(I)$, which can be weighted by the s_I if meaningful. (Compare the discussion in Chapter 4.)

2.3.5. *Applications*

Example 1. Estimation of average crop yields.

This survey was described on pages 35–38 where models for the population and the sample are also given.

The aim was to plan a balanced survey in which a farms, b fields per farm and c plots per field were to be sampled,[1] and it was meant that a, b and c were to be determined as efficiently as possible. The arithmetical mean $y...$ of all the observations was to be used as estimator. Its variance

[1] Farms and fields with the same unequal probabilities as for the two sets of data discussed below.

is $1/abc[\sigma^2 + c\sigma^2_{AB} + bc\sigma^2_A]$ where we have introduced $\sigma^2 = \sigma^2_\varepsilon + \sigma^2_{ABC}$. In order to determine optimum values of a, b and c one has to estimate the three components σ^2, σ^2_{AB} and σ^2_A.

Two sets of data were accessible. In the first a' farms, one field per farm and c' plots were selected. As this is a balanced situation we can use the ordinary analysis of variance between and within farms. The expected mean squares are

$$\mathrm{E}(\mathrm{MS}^{(1)}_A) = \sigma^2 + c'(\sigma^2_{AB} + \sigma^2_A)$$

$$\mathrm{E}(\mathrm{MS}^{(1)}_{ABC}) = \sigma^2$$

From this first set of data it is thus possible to obtain

$$\hat{\sigma}^2 = \mathrm{MS}^{(1)}_{ABC}$$

$$\hat{\sigma}^2_{AB} + \hat{\sigma}^2_A = \frac{\mathrm{MS}^{(1)}_A - \mathrm{MS}^{(1)}_{ABC}}{c'},$$

but not to separate the variances between farms and between fields.

In the second set a'' farms, $b_i = B_i$ fields from farm i and c_{ij} plots from field ij were sampled. Let us define the following mean squares between and within fields:

$$\mathrm{MS}^{(2)}_{AB} = \frac{1}{a''} \sum_{ij} p_{ij}(y_{ij.} - \hat{y}_{i..})^2$$

$$\mathrm{MS}^{(2)}_{ABC} = \frac{1}{a''} \sum_{ijk} p_{ij}(1 - p_{ij}) \frac{1}{c_{ij}(c_{ij} - 1)} (y_{ijk} - y_{ij.})^2$$

where p_{ij} are the probabilities (proportional to areas) with which the fields would have been sampled, if not all of them had been investigated, and where

$$\hat{y}_{i..} = \sum_j p_{ij} y_{ij.}$$

The expected mean squares are

$$\mathrm{E}(\mathrm{MS}^{(2)}_{AB}) = \sum_{IJ} p_I \cdot p_{IJ}(1 - p_{IJ}) \frac{1}{c_{IJ}} \cdot \sigma^2_{IJ} + \sigma^2_{AB}$$

$$\mathrm{E}(\mathrm{MS}^{(2)}_{ABC}) = \sum_{IJ} p_I \cdot p_{IJ}(1 - p_{IJ}) \frac{1}{c_{IJ}} \cdot \sigma^2_{IJ}$$

and thus $$\hat{\sigma}^2_{AB} = \mathrm{MS}^{(2)}_{AB} - \mathrm{MS}^{(2)}_{ABC}.$$

TABLE II: 7. *Number of surviving animals, means and variances of the intensity of fluorescence.*

Diets	Groups of animals		
	Controls	Diabetic rats without cataracts	Diabetic rats with cataracts
I	14	8	9
	64.4	86.7	100.9
	26.60	319.51	158.92
V	16	8	11
	65.0	91.2	100.3
	25.51	125.29	383.74

Now it it possible to determine a, b and c from

$$D^2(y\ldots) \approx \frac{1}{abc}\cdot \mathrm{MS}^{(1)}_{ABC} + \frac{1}{ab}\cdot(\mathrm{MS}^{(2)}_{AB} - \mathrm{MS}^{(2)}_{ABC}) + \frac{1}{a}\left\{\frac{\mathrm{MS}^{(1)}_{A} - \mathrm{MS}^{(1)}_{ABC}}{c'} - (\mathrm{MS}^{(2)}_{AB} - \mathrm{MS}^{(2)}_{ABC})\right\}$$

and some suitable cost function.

Example 6. Studies of the fluorescence of the eye lens in alloxan diabetic rats.

One of the experiments in this investigation was designed to study the influence of different fat diets on the fluorescence. At first the animals were divided into two groups, one of which was injected. Then each group was subdivided into two groups, one for each fat diet (I and V). After 6 weeks of these diets both lenses from each animal were investigated and the diabetic rats classified as with or without cataracts. In table II: 7 are given the number of surviving animals, means and variances of the intensity of fluorescence.

Using the modified unweighted analysis we obtain the following analysis of variance table:

Sources of variation	D.f.	MS
Diets	1	3.38
Groups of animals	2	670.86
Interaction	2	7.11
Within cells	60	20.92

The unmodified analysis gave the within-mean-square 16.26 (there is a misprint in (12)).

TABLE II: 8.

Mother no.	Father no. 1	2	3
	33	39	28
1	4.9	5.1	5.3
	0.6	0.3	0.6
	28	36	27
2	5.2	5.4	5.4
	0.4	0.3	0.7
	29	32	25
3	6.0	5.8	5.5
	0.4	0.3	0.7
	27	30	32
4	5.6	5.7	6.0
	0.3	0.9	0.5

As this is a Model-I-situation we are using the within-mean-square only when testing hypotheses.

Example 7. Studies of the number of cotyledon for pines.

This example has been mentioned shortly on page 63. Four mothers and three fathers were considered as randomly sampled from infinite populations. From each of the 12 resulting families 40 grains of seed were sampled (at random). The number of "surviving" plants and the mean values and variances for the numbers of cotyledon per plant are given for each family in table II: 8 (the figures in each cell are in order r_{ij} ; $y_{ij.}$; $\hat{s}^2_{ij}$).

As the populations are infinite we can use the model

$$y_{ijk} = \mu + a_i + b_j + c_{ij} + e_{ijk}$$

where a_i ; b_j ; c_{ij} and e_{ijk} are uncorrelated (but not independent) random variables with means = 0 and variances σ^2_A ; σ^2_B ; σ^2_{AB} and σ^2_e respectively.

The modified unweighted analysis gives the following table:

Sources of variation	D.f.	Mean squares	E(Mean squares)
Between mothers	3	0.3297	$\sigma^2 + \sigma^2_{AB} + 3\sigma^2_A$
Between fathers	2	0.0159	$\sigma^2 + \sigma^2_{AB} + 4\sigma^2_B$
Interaction	6	0.0481	$\sigma^2 + \sigma^2_{AB}$
Within families	354	0.0170[1]	σ^2

[1] This mean square and the corresponding one for the unmodified analysis are computed from the rounded off variances given above, whereas the other mean squares are quoted from (19).

What has here been denoted by σ^2 is something like

$$\underset{IJ}{E}\left\{\sigma^2_{e_{IJ}} \cdot E_{IJ}\left(\frac{1}{r_{IJ}}\right)\right\}$$

where $E_{IJ}(1/r_{IJ})$ is the expected value of $1/r_{IJ}$ given IJ and $\underset{IJ}{E}$ the expectation over all possible choices of families. Unbiased estimates of the variance components are

$$\hat{\sigma}^2_A = 0.0939$$

$$\hat{\sigma}^2_B = 0$$

$$\hat{\sigma}^2_{AB} = 0.0311$$

The unmodified unweighted analysis gives a within-mean-square of 0.0163, a very small deviation from the value in the modified analysis as could be expected when the r_{ij} and $\hat{s}^2_{ij}$ are not too variable.

2.4. Notes about the reliability of estimates

In the preceding paragraphs only point estimates for the variance components have been discussed. Of course it is important to know something about the variances of these estimates, especially when there are two or more analyses, giving different point estimates, to choose between. The methods to be found in the pertinent literature on this subject depend to a large extent on assuming normality and independently sampled effects in the models. An effort has been made in this paper to avoid making such assumptions.

The most general work has been done by Hooke (14), who has considered (special cases of) a balanced design for two factors of a model like those used in the foregoing discussion. His formulas are, however, very complicated. A review of methods for computing variances of estimates of variance components and for computing confidence limits can be found in Plackett (21).

CHAPTER 3

TESTING HYPOTHESES

3.1. Introduction

Analysis of variance is probably still used mostly as a means of testing. The classical assumptions are here reviewed for a two-factor-design. If the model is fixed we write it as follows:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$$

where $\Sigma_i \alpha_i = \Sigma_j \beta_j = \Sigma_i \gamma_{ij} = \Sigma_j \gamma_{ij} = 0$ and e_{ijk} are N.i.d. $(0; \sigma_e^2)$,[1] that is normally and independently distributed with means $= 0$ and a common variance σ_e^2. The sums of squares are then independently distributed as constants times χ^2 and the quotients between suitably chosen mean squares will follow the F-distribution, central F when the null hypothesis is true and non-central F otherwise.

The same elegance can be attained in a random model,

$$y_{ijk} = \mu + a_i + b_j + c_{ij} + e_{ijk},$$

where a_i; b_j; c_{ij} and e_{ijk} are independent random variables normally distributed with means $= 0$ and variances σ_A^2, σ_B^2, σ_{AB}^2 and σ_e^2 respectively. Here we have the problem that if the interaction term c_{ij} is "tied to" the terms a_i and b_j, the c_{ij} cannot at the same time be uncorrelated to a_i and b_j and normally distributed. Small deviations from normality are not important, moreover this problem will not trouble us much, as the primary interest in a random model is estimation.

Introducing the corresponding assumptions in a mixed model will show that the mean squares for the fixed factor and for interaction are not distributed as constants times χ^2 without using further, often unrealistic, assumptions (see Scheffé (24)). Therefore the usual F-test for the fixed factor is not exact.

Some problems concerning F-tests will be discussed in this chapter. In paragraph 3.2 two problems in normal-theory-models are described. The

[1] This can be relaxed to the case where the covariance matrix of the e_{ijk} is known up to a scalar factor. See for example Scheffé (24), chap. 1.5.

first problem concerns the mixed model mentioned above, and the second problem, relevant in random or mixed models with several factors, arises when we do not have two mean squares with the same expectation although the null hypothesis is true. The reason why normal-theory-tests are taken up is that they are generally tried in most situations, and then it is useful to be aware of all the difficulties.

Tests which try to remove most of the classical assumptions are discussed in paragraph 3.3. Permutation tests of a type investigated by, for example, Welch (32), Pitman (20) and Wilk (34) are given for some situations together with some numerical measures of how well they are approximated by the F-tests. It is also mentioned that t-tests on all possible pairs of treatments may sometimes be more robust than the F-test.

3.2. Problems in the classical normal theory

3.2.1. The mixed model, as introduced by Scheffé (23), can be described in the following way for a two-factor-design. $\mathcal{A}$ is the fixed factor with levels $i=1, 2 \ldots a$ och $\mathcal{B}$ the random factor with levels $J=1, 2 \ldots$. The "true values" for the factor combinations are $\mu(i, J)$. If $j=1, 2 \ldots b$ denotes the randomly chosen levels of factor $\mathcal{B}$ we have the random true values $u(i, j)$. When building up the model Scheffé uses the random vector $\mathbf{u}(j)=\{u(1, j); u(2, j); \ldots u(a, j)\}'$, with covariance matrix:

$$[\mathrm{Cov}\{u(i, j); u(i', j)\}] = [\sigma_{ii'}]$$

The model equation can be written

$$y_{ijk} = \mu + \alpha_i + b_j + c_{ij} + e_{ijk}$$

where

$$\begin{aligned}
&\mu = u(\cdot\,\cdot) = \mu(\cdot\,\cdot)\\
&\alpha_i = u(i\cdot) - u(\cdot\,\cdot) = \mu(i\cdot) - \mu(\cdot\,\cdot)\\
&b_j = u(\cdot\, j) - \mu(\cdot\,\cdot)\\
&c_{ij} = u(i, j) - \mu(i\cdot) - u(\cdot\, j) + \mu(\cdot\,\cdot)
\end{aligned}$$

Here the roman letter u stands for a random variable and, for example, $\mu(i\cdot)$ for a constant. As soon as u is averaged over the random factor $\mathcal{B}$ it is replaced by μ.

The side conditions are

$$\Sigma_i \alpha_i = 0; \quad \Sigma_i c_{ij} = 0 \text{ for all } j\text{'s}; \quad E(b_j) = 0; \quad E(c_{ij}) = 0 \text{ for all } i\text{'s};$$

and the variance components are defined as follows

$$\sigma_A^2 = \frac{1}{a-1}\sum_i \alpha_i^2$$

$$\sigma_B^2 = D^2(b_j)$$

$$\sigma_{AB}^2 = \frac{1}{a-1}\sum_i D^2(c_{ij})$$

With the help of the elements in the covariance matrix the last two can be written

$$\sigma_B^2 = \sigma_{..}$$

$$\sigma_{AB}^2 = \frac{1}{a-1}\sum_i (\sigma_{ii} - \sigma_{..})$$

Scheffé shows that the sums of squares for A and AB are not distributed as constants times χ^2 unless

$$\sigma_{ii'} = \varrho \cdot \sigma^2; \quad i \neq i'$$

and

$$\sigma_{ii} = \sigma^2$$

This means that the usual test of the fixed factor:

$$F = \frac{\mathrm{MS}_A}{\mathrm{MS}_{AB}}$$

is exact only when the variances within the levels of the fixed factor and the covariances between these levels are the same. If we look at the model equation we see that the variance for $\alpha_i + b_j + c_{ij}$ taken over all possible levels j (over the population of levels for factor $\mathcal{B}$) should be independent of i and that the covariances between $\alpha_i + b_j + c_{ij}$ and $\alpha_{i'} + b_j + c_{i'j}$ as well independent of i. These requirements are apparently fulfilled if $c_{ij} = 0$, i.e. if there is additivity, otherwise they are probably not met with in a practical problem. Box (5) has shown that this inexactness will result in overestimated significances (that is, the probability of rejecting a true null hypothesis is larger than nominal) if one uses the F-test.

An exact test for this situation based on Hotelling's T^2-statistic is described by Scheffé (24) for two factors and by Imhof in (15) for one fixed and two random factors.

The method requires however, very tedious numerical computations. It results for the two factor case in an F-test with $a-1$ and $b-a+1$ d.f.,

which makes it impossible to use when $a > b$ and gives it very low power when $b - a + 1$ is small.

Imhof (16) describes a quick but approximate method which as a rule has a higher power than the exact alternative. The quotient MS_A/MS_{AB} is used, but the degrees of freedom are "corrected" to

$$\bar{h}_{\min} \quad \text{resp.} \quad \bar{h}_{\min}(b-1),$$

where

$$\bar{h}_{\min} = \frac{a-1}{1+(a-2)(1-\bar{\varphi})^2}$$

and

$$\bar{\varphi} = \min\left(1; \frac{MS_{\text{Residual}}}{MS_{AB}}\right)$$

From the last relation it can be seen that, when the residual mean square is at least equal to the interaction mean square, the degrees of freedom are unchanged. Thus with no interaction the correction will be slight.

Example 3. Comparing washing-machines.

An application, introduced on pages 40–41, in which different makes of washing-machines were compared in several different trials (replications), has been thoroughly discussed by Seeger (26). The trials consisted of different batches of dirty garments. One wanted to generalize the results to some infinite population of trials. This population could not be expected to be as symmetrical as required for exactitude in the common F-test in a mixed model.

Numerical results for the experiment, in which 9 machines were compared in 15 trials and 4 pieces of garment measured per machine per trial, are presented below:

Sources of variation	D.f.	MS
Machines	8	383.60
Trials	14	271.76
Interaction	112	34.14
Residual	405	17.97

The usual F-test for machines gives $F = 383.60/34.14 = 11.24$ which is highly significant as the critical value with level 0.05% is ~ 3.9. Using Hotelling's T^2 one obtains $F = 24.44$ with 8 and 7 degrees of freedom. The corresponding critical value at level 0.05% is 18.2 which means that the

latter method brought us somewhat nearer the 0.05% limit. The difference between the methods becomes more accentuated when one wants to make individual comparisons between the machines.[1] To show this we give the mean values for the 9 machines *A–I*.

A 89.1	D 84.4	G 85.2
B 89.5	E 88.0	H 82.4
C 84.9	F 85.6	I 83.3

If the *S*-method connected to the usual *F*-test is used for comparisons between all possible pairs of machines the results are as follows (level 5%):

A is significantly better than D, H, I
B is significantly better than D, H, I, C
E is significantly better than H, I

The *S*-method built on Hotelling's T^2 gives no significant paired comparison.

With Imhof's approximate method the degrees of freedom are "corrected" from 8 and 112 to 3.11 and 43.54. The critical value at the level 0.05% is then ~ 7. For the corresponding paired comparisons (*S*-method at level 5%) one obtains:

A is significantly better than H, I
B is significantly better than H, I, D
E is significantly better than H

In (26) there are also given numerical comparisons between the usual *F*-test and that based on Hotelling's T^2 for several examples. The difference is as a rule very large, especially of course when $b-a$ is small.

It seems to be unclear whether it is worth while to use something else than the usual *F*-quotient for testing the fixed factor. Some sampling experiments will be presented for a mixed model in part IV. There most of the classical assumptions, including that of symmetry discussed above, are avoided, and still the usual *F*-test seems to be of practical use.

3.2.2. Let us now consider a sampling case with three crossed factors $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$. The populations of levels are assumed to be infinite for $\mathcal{A}$ and $\mathcal{B}$ and exhausted ($c=C$) for the factor $\mathcal{C}$. It is then easy to use the rules on pages 46–47 and write out the E(MS)'s for a balanced situation with *abcr* observations.

---

[1] This problem will be dealt with in part III, where the below mentioned *S*-method among others, for making multiple comparisons, is described.

$$
\begin{aligned}
\mathrm{E}(\mathrm{MS}_A) &= \sigma^2 + rc\sigma^2_{AB} + rbc\sigma^2_A \\
\mathrm{E}(\mathrm{MS}_B) &= \sigma^2 + rc\sigma^2_{AB} + rac\sigma^2_B \\
\mathrm{E}(\mathrm{MS}_{AB}) &= \sigma^2 + rc\sigma^2_{AB} \\
\mathrm{E}(\mathrm{MS}_C) &= \sigma^2 + r\sigma^2_{ABC} + ra\sigma^2_{BC} + rb\sigma^2_{AC} + rab\sigma^2_C \\
\mathrm{E}(\mathrm{MS}_{AC}) &= \sigma^2 + r\sigma^2_{ABC} + rb\sigma^2_{AC} \\
\mathrm{E}(\mathrm{MS}_{BC}) &= \sigma^2 + r\sigma^2_{ABC} + ra\sigma^2_{BC} \\
\mathrm{E}(\mathrm{MS}_{ABC}) &= \sigma^2 + r\sigma^2_{ABC} \\
\mathrm{E}(\mathrm{MS}_E) &= \sigma^2
\end{aligned}
$$

If we make all the classical assumptions we have a mixed model with the two factors $\mathcal{A}$ and $\mathcal{B}$ random and the factor $\mathcal{C}$ fixed. For the reason discussed above tests for BC and AC are not exact, for C there is not even a mean square to be used as a comparison. We see that it is probably true that $\mathrm{E}(\mathrm{MS}_C) > \mathrm{E}(\mathrm{MS}_{AC})$ and $> \mathrm{E}(\mathrm{MS}_{BC})$, assuming the null hypothesis H_0: $\sigma^2_C = 0$.

One possible test would be to use

$$
F = \frac{\mathrm{MS}_C}{\mathrm{MS}_{AC} + \mathrm{MS}_{BC} - \mathrm{MS}_{ABC}} \tag{II.6}
$$

with the degrees of freedom according to Satterthwaite (22)

$$
f_C \quad \text{resp.} \quad \frac{(\mathrm{MS}_{AC} + \mathrm{MS}_{BC} - \mathrm{MS}_{ABC})^2}{\mathrm{MS}^2_{AC}/f_{AC} + \mathrm{MS}^2_{BC}/f_{BC} + \mathrm{MS}^2_{ABC}/f_{ABC}}
$$

The latter is determined in order to approximate the distribution of the denominator by a χ^2-distribution with unchanged variance, applying the fact that MS_{AC}, MS_{BC} and MS_{ABC} are in principle distributed as χ^2 is. As has been said above none of the four mean squares used in (II.6) is distributed as χ^2 is, unless some symmetry restrictions are fulfilled. Thus a large number of approximations have been introduced.

Cochran (6) has pointed out that the distribution of a linear combination of mean squares (each distributed as a constant times χ^2) is better approximated by a χ^2-distribution when all the coefficients in the linear combination are positive than when some of them are negative. He therefore proposes to use a linear combination with positive coefficients in both numerator and denominator. Some numerical computations in the case where only three mean squares are involved showed an excellent approximation for the level 5%. In the example above and according to Cochran we could have used:

$$
F = \frac{\mathrm{MS}_C + \mathrm{MS}_{ABC}}{\mathrm{MS}_{AC} + \mathrm{MS}_{BC}}
$$

where the numerator has the same expectation as the denominator when the null hypothesis is true but a larger expectation when the alternative hypothesis is true. The method used in (II.6) will however sometimes give us an exactly χ^2-distributed variable in the numerator, while Cochran's method always gives approximations in both numerator and denominator.

Box (4), using known positive coefficients for χ^2-distributed variables, gives numerical results of the true risks for type-I-errors which are very close to the nominal 5, 10, 25 and 50%. However, in analysis of variance the coefficients are $c \cdot (\sigma^2/f)$[1] and have to be estimated.

There seems to be no directly convincing evidence in the literature on this subject in favour of any of the two methods.

3.3. General models with few assumptions

3.3.1. *Illustration of a permutation test*

R. A. Fisher 1925 (10) was probably the first to use the so-called permutation tests in order to avoid the usual assumptions of normality etc. The idea behind such tests can be illustrated by the following small example:

Three types of wheat, A_1, A_2 and A_3 are compared in three blocks, each block with three plots. The treatments are randomized over the plots in each block. Let us assume that the results of the experiment are the following, where the figures are some measures of yield:

Block no.	Plot no. 1	Plot no. 2	Plot no. 3
1	3 A_1	4 A_3	6 A_2
2	8 A_2	7 A_3	11 A_1
3	6 A_2	3 A_1	5 A_3

If now the null hypothesis, which says that on the average the treatments will give the same yield in every plot is true, we can consider the obtained yields as reflecting differences between plots only. Ignoring measurement errors the results would have been the same for any outcome of the randomization. Thus one can determine the distribution of all possible results and see if the one obtained hints that the null hypothesis is not true.

[1] A SS_Q is (except for fixed Q in a mixed model) distributed as $E(MS_Q) \cdot \chi^2$.

As can be understood from this it is not possible to test the hypothesis that the treatments are equal on the average over all plots (H_0: $\sigma_A^2 = 0$), but only that they are equal for every plot (H_0: $\sigma_A^2 = 0$ and $\sigma_{AB}^2 = 0$ and $\sigma_{ABP}^2 = 0$).

Schematically a permutation test is executed as follows:

1) One chooses a level of significance, α;
2) one computes the value of some suitable statistic, say F, for the obtained result;
3) one considers all possible permutations of the observations (3! for every block in the example above) and computes F for every one;
4) one computes the proportion of F-values that exceed the actually obtained F and rejects the hypothesis if this proportion is $\leqslant \alpha$.

3.3.2. *Normal approximations*

For larger designs this is only theoretically possible, even with the help of the fastest electronic computer. In one of the experiments with washing-machines discussed in (26) it was required to consider $(9!)^{59}$ different permutations. Therefore it is natural to look for some approximative method. For many situations the Model I normal-theory-test has proved to be a good approximation to the permutation test. Welch (32) and Pitman (20) have studied the case of one factor with completely randomized, randomized block[1] and Latin square designs for completely exhausted populations and no technical errors. They test the hypothesis that the treatment levels are identical with the help of

$$U = \frac{SS_A}{SS_A + SS_{Res.}} = \frac{\text{sum of squares for treatments}}{\text{sum of squares for treatments} + \text{residual}},$$

where the denominator is constant over all permutations.

In normal theory U follows the β-distribution and corresponds directly to

$$F = \frac{MS_A}{MS_{Res.}} \quad \text{(See Scheffé (24) page 322.)}$$

Welch and Pitman show for what values of the observations the first four moments of U under normal-theory-model and permutation-model are equal. They also introduce a correction factor to the numbers of degrees of freedom for the normal-theory-test, making the first two moments equal.

---

[1] Only $r = 1$ plot per treatment and block.

Wilk (34) generalizes these results to the case where the treatments are replicated $r > 1$ times per block. He uses the test criterion

$$\frac{SS_A + SS_{AB}}{SS_A + SS_{AB} + SS_{Res.}}$$

Scheffé (24), by considering conditioning, shows how technical errors can be introduced.

3.3.3. *An application*

Here we will treat a rather complicated case from the study of washing-machines. Further details can be found in (26).

Example 3. Comparing washing-machines.

In this example 9 machines were compared in 15 trials. For every trial the pieces of garment were grouped according to how dirty they were and then randomized among machines. Four blocks (groups), each with one garment per machine were used. The machine-factor was denoted by $\mathcal{A}$, trials by $\mathcal{B}_1$, blocks by $\mathcal{B}_1\,\mathcal{B}_2$ and plots (pieces of garment) by $\mathcal{B}_1\,\mathcal{B}_2\,\mathcal{P}$. By considering fixed plots and blocks we avoid the problem mentioned on page 41, caused by the grouping of the sampled pieces of garment. The following population model was used

$$\mu(IJMK) = \mu + \alpha(I) + \beta_1(J) + \lambda(IJ) + \beta_2(JM) + \delta(IJM) + \varepsilon(JMK) + \eta(IJMK)$$

in which the easily recognizable effects are: α for treatments, β_1 for trials, λ for treatment-trial interaction, β_2 for blocks, δ for treatment-block interaction, ε for plots, η for treatment-plot interaction.

This case with two block systems is not covered by the discussion in paragraph 2.2, but the expected mean squares are given in (26) for fixed treatments, blocks and plots and an infinite number of potential trials:

$$E(MS_A) = \sigma_e^2 + \sigma^2_{AB_1B_2P} + \sigma^2_{B_1B_2P} + b_2\left[\sigma^2_{AB_1} - \frac{1}{B_2}\cdot\frac{1}{A}\frac{A}{A-1}\cdot\sigma^2_{AB_1B_2P}\right] + b_2b_1\sigma^2_A$$

$$E(MS_{B_1}) = \sigma_e^2 + \sigma^2_{AB_1B_2P} + b_2a\sigma^2_{B_1}$$

$$E(MS_{B_1B_2}) = \sigma_e^2 + \sigma^2_{AB_1B_2P} + a\sigma^2_{B_1B_2}$$

$$E(MS_{AB_1}) = \sigma_e^2 + \sigma^2_{AB_1B_2P} + \sigma^2_{B_1B_2P} + b_2\left[\sigma^2_{AB_1} - \frac{1}{B_2}\cdot\frac{1}{A}\cdot\frac{A}{A-1}\cdot\sigma^2_{AB_1B_2P}\right]$$

$$E(MS_{AB_1B_2}) = \sigma_e^2 + \sigma^2_{AB_1B_2P} + \sigma^2_{B_1B_2P} + \left[\sigma^2_{AB_1B_2} - \frac{1}{A}\cdot\frac{A}{A-1}\cdot\sigma^2_{AB_1B_2P}\right]$$

Even if the most interesting hypothesis is $H_0 : \sigma_A^2 = 0$ it has to be replaced by $H_0 : \sigma_A^2 = \sigma_{AB_1}^2 = \sigma_{AB_1B_2}^2 = \sigma_{AB_1B_2P}^2 = 0$ (the treatments being identical in every plot), if we want to use a permutation test. In order to be able to approximate the permutation test with a normal-theory-test, it was necessary to assume $\sigma_{AB_1B_2}^2 = 0$, and test the hypothesis $H_0 : \sigma_A^2 = \sigma_{AB_1}^2 = \sigma_{AB_1B_2P}^2 = 0$.

The test criterion

$$\frac{SS_A + SS_{AB_1}}{SS_A + SS_{AB_1} + SS_{AB_1B_2}}$$

was used and denoted by U in the normal-theory-model and U' in the permutation model.[1] When H_0 is true U follows a β-distribution with $f_1 = b_1(a-1)$ and $f_2 = b_1(a-1)(b_2-1)$ degrees of freedom. To reject the null hypothesis if $U > U_0$ is equivalent to rejecting it if $F > F_0$,

where

$$F = \frac{(1/f_1) \cdot (SS_A + SS_{AB_1})}{(1/f_2) \cdot SS_{AB_1B_2}}$$

If we denote numerator and denominator in this expression by N and D we obtain in the permutation model

$$E(N) = E(D) = \sigma_e^2 + \sigma_{B_1B_2P}^2 \text{ when } H_0 \text{ is true}$$

and

$$E(N) = \sigma_e^2 + \sigma_{B_1B_2P}^2 + \frac{a-2}{a-1}\sigma_{AB_1B_2P}^2 + b_2\sigma_{AB_1}^2 + b_2\sigma_A^2$$

$$E(D) = \sigma_e^2 + \sigma_{B_1B_2P}^2 + \frac{a-2}{a-1}\sigma_{AB_1B_2P}^2$$

when it is not.

As soon as any of $\sigma_{AB_1}^2$ and σ_A^2 is >0, there is a tendency for N to be $> D$. Increasing $\sigma_{AB_1B_2P}^2$ will decrease the power of the test.

Next the first two moments for U and U' will be compared under H_0. If we begin by disregarding the technical errors, the statistical model can be written:

$$y_{IJM} = \mu + \sum_J b_j(J)\,\beta_1(J) + \sum_J b_j(J)\,\beta_2(JM) + \sum_{JK} b_j(J)\,d_{IJM}(K)\,\varepsilon(JMK)$$

when H_0 is true. Here the indicator variables are defined as follows:

$$b_j(J) = \begin{cases} 1 \text{ if trial } j \text{ in the sample corresponds to trial } J \text{ in the population} \\ 0 \text{ otherwise} \end{cases}$$

[1] This criterion is not necessarily monotonically increasing with increasing deviations from H_0, unless $\sigma_{AB_1B_2}^2 = 0$.

$$d_{IjM}(K)=\begin{cases}1 \text{ if plot } K \text{ within } jM \text{ is randomized to treatment (machine) } I\\ 0 \text{ otherwise}\end{cases}$$

Capital letters I and M have been used instead of i and m to indicate that the factors $\mathcal{A}$ and $\mathcal{B}_2$ are fixed.

In the following derivations we consider the b_1 chosen trials as given, the generalization to the infinite population of trials will be discussed later.

Mean and variance for U

Kempthorne (17) has shown that

$$E(U)=\frac{f_1}{f_1+f_2}=\frac{1}{b_2}$$

$$D^2(U)=\frac{2f_1f_2}{(f_1+f_2)^2(f_1+f_2+2)}=\frac{2(b_2-1)}{b_2^2\{b_1b_2(a-1)+2\}}$$

Mean for U'

Taking into account that the denominator in the test criterion is constant over the permutation one obtains

$$E(U')=\frac{E(\mathrm{SS}_A+\mathrm{SS}_{AB_1})}{\mathrm{SS}_A+\mathrm{SS}_{AB_1}+\mathrm{SS}_{AB_1B_2}}$$

$$=\frac{\{(a-1)+(a-1)(b_1-1)\}\sigma^2_{B_1B_2P}}{\{(a-1)+(a-1)(b_1-1)+(a-1)b_1(b_2-1)\}\sigma^2_{B_1B_2P}}=\frac{1}{b_2}$$

Thus the means for U and U' are always equal.

Variance for U'

The total sum of squares, which is a constant, can be partitioned into sums of squares between and within blocks:

$$\sum_{IjM}(y_{IjM}-y_{...})^2=\sum_{IjM}(y_{IjM}-y_{.jM})^2+\sum_{IjM}(y_{.jM}-y_{...})^2$$

The denominator in U' is just the variation within blocks and can thus be written

$$\mathrm{SS}_A+\mathrm{SS}_{AB_1}+\mathrm{SS}_{AB_1B_2}=\sum_{IjM}(y_{IjM}-y_{.jM})^2$$

The numerator is

$$\mathrm{SS}_A+\mathrm{SS}_{AB_1}=b_2\sum_{Ij}(y_{I..}-y_{...})^2+b_2\sum_{Ij}(y_{Ij.}-y_{I..}-y_{.j.}+y_{...})^2=b_2\sum_{Ij}(y_{Ij.}-y_{.j.})^2$$

$$=\frac{1}{b_2}\sum_{Ij}\left\{\sum_{M}(y_{IjM}-y_{.jM})\right\}^2=\frac{1}{b_2}\sum_{IjM}(y_{IjM}-y_{.jM})^2$$

$$+\frac{1}{b_2}\sum_{Ij\,M\neq M'}(y_{IjM}-y_{.jM})(y_{IjM'}-y_{.jM'})$$

With help of the statistical model we obtain

$$y_{IjM} - y_{.jM} = \sum_{K} d_{IjM}(K)\,\varepsilon(jMK)$$

$$\sum_{IjM} (y_{IjM} - y_{.jM})^2 = \sum_{IjMK} d_{IjM}(K)\,\varepsilon^2(jMK) = \sum_{jMK} \varepsilon^2(jMK)$$

$$E\{b_2^2(\mathrm{SS}_A + \mathrm{SS}_{AB_1})^2\} = \Big\{ \sum_{jMK} \varepsilon^2(jMK) \Big\}^2 + E\Big\{ \sum_{\substack{IjKK' \\ M \neq M'}} d_{IjM}(K) d_{IjM'}(K')\,\varepsilon(jMK)\,\varepsilon(jM'K') \Big\}^2 .$$

After some computations the following result comes out:

$$E(\mathrm{SS}_A + \mathrm{SS}_{AB_1})^2 = \frac{1}{b_2^2} \Big\{ \sum_{jMK} \varepsilon^2(jMK) \Big\}^2 + \frac{2}{b_2^2(a-1)} \sum_{\substack{jKK' \\ M \neq M'}} \varepsilon^2(jMK)\,\varepsilon^2(jM'K')$$

$$D^2(U') = E(U'^2) - \{E(U')\}^2 = \frac{E(\mathrm{SS}_A + \mathrm{SS}_{AB_1})^2}{(\mathrm{SS}_A + \mathrm{SS}_{AB_1} + \mathrm{SS}_{AB_1B_2})^2} - \frac{1}{b_2^2}$$

$$= \frac{2}{b_2^2(a-1)} \cdot \frac{\sum_{\substack{jKK' \\ M \neq M'}} \varepsilon^2(jMK)\,\varepsilon^2(jM'K')}{\Big\{ \sum_{jMK} \varepsilon^2(jMK) \Big\}^2}$$

This variance is to be compared with

$$D^2(U) = \frac{2(b_2 - 1)}{b_2^2\{b_1 b_2(a-1) + 2\}}$$

Special case (i)

If we assume that $\sum_{MK} \varepsilon^2(jMK) = \sum_{MK} \varepsilon^2(MK)$, that is that the plot errors are equal within trials, we obtain:

$$D^2(U') = \frac{2}{b_2^2 b_1(a-1)} \cdot \frac{(1/b_1) \sum_{jKK'M \neq M'} \varepsilon^2(MK)\,\varepsilon^2(M'K')}{\Big\{ \sum_{MK} \varepsilon^2(MK) \Big\}^2}$$

to be compared with

$$D^2(U) = \frac{2}{b_2^2 b_1(a-1)\{b_2/(b_2-1) + 2/b_1(b_2-1)(a-1)\}}$$

Special case (ii)

Here we assume that the plot-errors are equal for every block, which gives:

$$\Big\{\sum_{jMK}\varepsilon^2(jMK)\Big\}^2 = b_1^2 b_2^2 (a-1)^2 \cdot \sigma^4_{B_1B_2P}$$

$$\sum_{jKK'M\neq M'}\varepsilon^2(jMK)\,\varepsilon^2(jM'K') = b_1 b_2 (b_2-1)(a-1)^2 \cdot \sigma^4_{B_1B_2P}$$

and

$$D^2(U') = \frac{2(b_2-1)}{b_2^3 b_1 (a-1)}; \quad D^2(U) = \frac{2(b_2-1)}{b_2^3 b_1 (a-1)} \cdot \frac{1}{\{1+2/b_1 b_2 (a-1)\}}$$

For the discussed application with $b_1 = 15$, $b_2 = 4$, $a-1 = 8$ is obtained

$$D^2(U') = 0.00078125; \quad D^2(U) = 0.00077798.$$

Thus $D^2(U')$ is slightly greater than $D^2(U)$ for this special case, which means that the normal-theory-test will overestimate the significance to an unimportant degree compared to the permutation-test.

A numerical value for the general expression

$$D^2(U') = \frac{2}{b_2^2(a-1)} \cdot \frac{\sum_{jKK'M\neq M'}\varepsilon^2(jMK)\,\varepsilon^2(jM'K')}{\Big\{\sum_{jMK}\varepsilon^2(jMK)\Big\}^2} = \frac{2}{b_2^2(a-1)} \cdot V$$

can be computed from the experiment. It was shown on page 92 that

$$\sum_{jMK}\varepsilon^2(jMK) = \sum_{IjM}(y_{IjM} - y_{.jM})^2 = \mathrm{SS}_A + \mathrm{SS}_{AB_1} + \mathrm{SS}_{AB_1B_2}$$

and thus

$$\sum_{jKK'M\neq M'}\varepsilon^2(jMK)\,\varepsilon^2(jM'K') = \sum_{jM\neq M'}\Big\{\sum_I (y_{IjM} - y_{.jM})^2 \cdot \sum_I (y_{IjM'} - y_{.jM'})^2\Big\}$$

For our numerical example the result is $V = 0.058576$ and $D^2(U') = 0.0009153$, thus a somewhat larger deviation from $D^2(U)$ than above.

It is easily shown that

$$0 \leqslant V \leqslant \frac{b_2-1}{b_2}$$

that is

$$0 \leqslant D^2(U') \leqslant \frac{2(b_2-1)}{b_2^3(a-1)} = 0.011719$$

Thus theoretically $D^2(U')$ may be much larger than $D^2(U)$, about b_1 times larger. For one system of randomized blocks Welch (32) has shown that the second moment in the permutation model, when at a maximum, can only be slightly larger than that of the normal-theory. Welch's result is in this paper obtained as a special case for $b_1 = 1$.

Equalizing $D^2(U)$ and $D^2(U')$

Using the normal approximation it is apparently possible to make both large overestimations and large underestimations. But as we can compute $E(U')$ and $D^2(U')$ given the b_1 blocks we can modify the degrees of freedom for the β-variable in order to equalize the first two moments.

The system of equations:

$$E(U) = \frac{f_1}{f_1 + f_2} = E(U')$$

$$D^2(U) = \frac{2f_1 f_2}{(f_1 + f_2)^2 (f_1 + f_2 + 2)} = D^2(U')$$

gives

$$f_1 = \Phi b_1 (a - 1)$$

$$f_2 = \Phi b_1 (b_2 - 1)(a - 1)$$

with

$$\Phi = \frac{1}{b_1 b_2} \left(\frac{b_2 - 1}{V} - \frac{2}{a - 1} \right)$$

(As $V \leqslant (b_2 - 1)/b_2$ we have $\Phi \geqslant 1/b_1 - 2/b_1 b_2 (a - 1) = 0.0625$.) Thus it would be possible in applications to compute Φ from the observations and correct the degrees of freedom in the normal-theory test. This correction factor has been given by Scheffé (24) for one system of randomized blocks.

Generalizations.—For the permutation test the risk of an error of the first kind is $\leqslant \alpha$ for every choice of b_1 trials. The unconditional probability of an error of the first kind is then also $\leqslant \alpha$. If the normal approximation gives a probability around α, given the b_1 trials, it will do so even unconditionally. However, it is not possible to check the accuracy for anything but the b_1 trials actually chosen.

In the same way technical errors independent of the treatments can be introduced. (See Scheffé (24).)

Another possibility would be to imagine the permutation test executed so that all combinations of b_1 trials from B_1 are chosen, and that all permutations are considered for each such choice. If we can imagine a test with $(9!)^{59}$ different permutations, we can as well imagine one with infinitely

many and see what the correspondence to the normal-theory-test will be.

In order to have $SS_A + SS_{AB_1} + SS_{AB_1B_2}$ constant we must assume that the plot-errors are equal from trial to trial, that is

$$\frac{1}{b_2(a-1)} \sum_{MK} \varepsilon^2(JMK) = \sigma^2_{B_1B_2P_J} = \sigma^2_{B_1B_2P} \text{ for all } J's.$$

Then we obtain

$$D^2(U') = \frac{2}{b_2^2 b_1(a-1)} \cdot \frac{(1/B_1) \sum_{JKK'M \neq M'} \varepsilon^2(JMK)\, \varepsilon^2(JM'K')}{\{\sum_{MK} \varepsilon^2(JMK)\}^2}$$

which cannot now be computed from the observations. Introducing

$$\sum_{K} \varepsilon^2(JMK) = (a-1)\, \sigma^2_{B_1B_2P;\, JM}$$

we obtain

$$D^2(U') = \frac{2}{b_2^2 b_1(a-1)} \left(1 - \frac{(1/B_1) \sum_{JM} \sigma^4_{B_1B_2P;\, JM}}{b_2^2 \sigma^4_{B_1B_2P}}\right).$$

Scheffé's result for randomized blocks in the fixed case is with our notations:

$$D^2(U') = \frac{2}{b^2(a-1)} \left(1 - \frac{\sum_{j} \sigma^4_{BP_j}}{b^2 \sigma^4_{BP}}\right)$$

The number of blocks is b and they are indicated by j.

Numerical results

In the actual application the analysis can be summarized as in table II: 9. (Correction factor $\Phi = 0.85$.)

TABLE II: 9. *Analysis of variance for an experiment comparing washing-machines (factor $\mathcal{A}$). The experimental units were divided in trials, $\mathcal{B}_1$, and blocks within trials, $\mathcal{B}_1\mathcal{B}_2$.*

Sources of variation	D.f.	MS	E(MS)
Machines (A)	8	383.60	$\sigma_e^2 + \frac{7}{8}\sigma^2_{AB_1B_2P} + \sigma^2_{B_1B_2P} + 4\sigma^2_{AB_1} + 60\sigma^2_A$
Trials (B)	14	271.76	$\sigma_e^2 + \sigma^2_{AB_1B_2P} + 36\sigma^2_{B_1}$
Blocks within trials (B_1B_2)	45	26.83	$\sigma_e^2 + \sigma^2_{AB_1B_2P} + 9\sigma^2_{B_1B_2}$
Machines · Trials (AB_1)	112	34.14	$\sigma_e^2 + \frac{7}{8}\sigma^2_{AB_1B_2P} + \sigma^2_{B_1B_2P} + 4\sigma^2_{AB_1}$
Machines · Blocks (AB_1B_2)	360	16.86	$\sigma_e^2 + \frac{7}{8}\sigma^2_{AB_1B_2P} + \sigma^2_{B_1B_2P} + \sigma^2_{AB_1B_2}$

TABLE II: 10. *Computed corrections to the degrees of freedom for 9 experiments with washing-machines.*

a = the number of machines, b = the number of trials, r = the number of replications, Φ = correction factor.

Experiment no.	a	b	r	Φ
1	7	9	6	0.819
2	7	10	3	0.876
3	7	10	3	1.004
4	9	15	6	0.751
5	9	15	3	0.886
6	9	15	3	0.981
7	9	10	6	0.899
8	9	10	3	0.972
9	9	10	3	0.986

Thus the test for machines is

$$F = \frac{1/120(8 \cdot 383.60 + 112 \cdot 34.14)}{16.86} = 3.41$$

which is highly significant at level 0.05% when the degrees of freedom are $0.85 \cdot 120 = 102$ and $0.85 \cdot 360 = 306$.

From table II: 9 one obtains

$$\hat{\sigma}_A^2 = 5.82$$

$$\hat{\sigma}_{AB_1}^2 = 4.32 + \tfrac{1}{4}\sigma_{AB_1B_2}^2,$$

which indicates that average differences between machines and interaction between machines and trials are equally responsible for the significant result.

3.3.4. *Numerical values for the correction factor.*

The correction factor Φ has been computed for 9 experiments in the study of washing-machines. In these experiments no blocking was made within trials. The results together with the number of levels and replications are shown in table II: 10.

When $\Phi < 1$ a normal-theory-test with uncorrected degrees of freedom will overestimate the significance compared to the permutation test.

3.3.5. *t-tests for pairs*

In the permutation test normality and equal variances are not assumed, only the randomization is utilized. Unequal variances hardly affect the

t-test in a balanced situation (if the samples are matched the t-test is not affected at all), and deviations from normality have rather unimportant effects unless the samples are very small. Therefore it would be possible to replace the F-test by all possible t-tests for pairs between treatments, if the individual significance levels are heightened (lowered) to account for the mass-significance. This problem will be thoroughly discussed in part III.

3.4. Conclusions

Some problems encountered in testing hypotheses are discussed in this chapter. First the problem of testing fixed factors and fixed or mixed interactions in mixed models was taken up. The usual F-tests for these components are not exact even if all the classical assumptions are satisfied, unless some symmetry requirements are fulfilled. An exact, but computationally tedious method with bad power based on Hotelling's T^2 is mentioned, and a rapid approximation with better power by Imhof (16) described. However, not much is known about how inexact the common F-test is. Some sampling experiments presented in part IV indicate that the F-test works reasonably well in most practical cases.

Another problem, that may arise in experiments with many factors, is that no single mean square is suitable as the denominator in the F-quotient. It would then be possible to use a linear function of mean squares in the denominator and to approximate the degrees of freedom like Satterthwaite (22). Cochran (6) has suggested the use of linear functions both in the numerator and the denominator so as to keep the coefficients for the mean squares positive, which will give better approximations to χ^2. It is unclear what is to be gained by this process in practice. The first method, being the most natural, is therefore to be preferred.

Permutation tests have also been discussed and detailed comparisons with a normal approximation are given for a practical example. It is mentioned how, using the observations, the degrees of freedom in the normal-theory-test can be modified to give a better approximation. The correction factor is computed for ten experiments to give an idea of its range for a certain type of data. Such permutation tests, and approximations to them, are, however, complicated both to execute and to interpret. They are difficult to interpret, because it is only possible to test null hypotheses in which several components are zero (see the example above.)

Summing up the foregoing, it will always be recommendable to use an ordinary F-test with the mean square connected to the null hypothesis in the numerator, and the mean square (linear combination of mean squares)

with the same expectation as the numerator under H_0, in the denominator. This will give a good enough approximation, unless the populations are very extraordinary (see again the sampling experiments in part IV.) The character (extraordinary or not) of the populations could well stand investigation before the experiment is performed.

In unbalanced designs the recommendations are the same, using the modified unweighted analysis. It is however more dubious that the approximation will be adequate. Yet, tests of hypotheses are not very important in practical work. They should be interpreted as giving loose indications rather then exact probability statements.

In the discussion above it is everywhere assumed that only independent observations are used.

CHAPTER 4

INDIVIDUAL COMPARISONS

4.1. Introduction

In the preceding two chapters we have been discussing estimation of variance components, for example in the fixed case of $1/(A-1)\Sigma_I \alpha^2(I)$, and hypothesis testing, for example of H_0: $\Sigma_I \alpha^2(I)=0$. Now we turn to the estimation of linear combinations $\Sigma_I c_I \alpha(I)$ in the effects of a fixed factor. With a fixed factor we will put $i=I$, so that α_i and $\alpha(I)$ can be used interchangeably.[1] The linear combination

$$\psi = \sum_I c_I \alpha(I)$$

is called a contrast if $\Sigma_I c_I = 0$. Most individual comparisons used in practice are contrasts, for example

$$\alpha(1)+\alpha(2)-2\cdot\alpha(3)$$

$$\alpha(1)-\alpha(2)$$

in which the coefficients c_I are $+1$, $+1$, -2 and $+1-1$, 0 respectively.

Two contrasts

$$\psi_1 = \sum_I c_{1I} \alpha(I)$$

and

$$\psi_2 = \sum_I c_{2I} \alpha(I)$$

are said to be orthogonal if the product sum of coefficients vanishes, that is $\Sigma_I c_{1I} \cdot c_{2I} = 0$. The two contrasts above are orthogonal as $(+1)(+1)+(+1)(-1)+(-2)(0)=0$. For a factor with a levels, $a-1$ mutually orthogonal contrasts can be found, one for each degree of freedom. In fact several such sets of $a-1$ orthogonal contrasts can be constructed. The nature of the levels can often determine which set is of most interest. The use of orthogonal contrasts makes the analysis very elegant. To each

[1] In pertinent literature exhaustive random sampling has sometimes been considered (Cornfield-Tukey (8)), that is $\alpha_i = \Sigma_{I=1}^{a=A} a_i(I)\alpha(I)$. This is mathematically convenient but hardly realistic.

contrast there corresponds one sum of squares and one degree of freedom, both of which add up to the totals for the factor. The sums of squares for orthogonal contrasts are thus independent.

The contrasts to be estimated are called population contrasts and the estimators, for example

$$\hat{\psi} = \sum_I c_I y_{I.}$$

are called sample contrasts. The orthogonality condition for sample contrasts in unbalanced situations is

$$\sum_I c_{1I} \cdot c_{2I} \cdot \frac{1}{r_I} = 0$$

where r_I is the number of observations making up the mean $y_{I.}$.

Important problems here are to determine which contrasts are relevant and how to estimate them. These problems are treated in paragraph 4.2. In the next paragraph variances and estimated variances are derived for different situations.

4.2. Choice of contrasts

As has been said, contrasts, and even orthogonal contrasts, can be chosen in many ways. Of course, one should in practice look for contrasts of interest, not necessarily orthogonal.

What contrasts are then of interest? Paired comparisons, as $\alpha_i - \alpha_i$ often interest us. They are always estimated by $y_{i.} - y_{i.'}$. Contrasts of the type

$$\tfrac{1}{2}(\alpha_1 + \alpha_2) - \tfrac{1}{3}(\alpha_3 + \alpha_4 + \alpha_5)$$

are also frequent. They are estimated by $\frac{1}{2}(y_{1.} + y_{2.}) - \frac{1}{3}(y_{3.} + y_{4.} + y_{5.})$ and sometimes, in the unbalanced case, by

$$\frac{r_1 y_{1.} + r_2 y_{2.}}{r_1 + r_2} - \frac{r_3 y_{3.} + r_4 y_{4.} + r_5 y_{5.}}{r_3 + r_4 + r_5}.$$

Such contrasts should not be used indiscriminately. If there are 5 different populations 1, 2, ... 5, is there any meaning in using the contrast above even if 1, 2 and 3, 4, 5 are related? Not unless one population had been stratified into the parts 1 and 2 and another into 3, 4 and 5. Then the relevant population contrast would be:

$$\frac{R_1\alpha_1 + R_2\alpha_2}{R_1 + R_2} - \frac{R_3\alpha_3 + R_4\alpha_4 + R_5\alpha_5}{R_3 + R_4 + R_5}$$

if R_i denotes the strata sizes. If r_i are proportional to R_i the unbalanced sample contrast above is an unbiased estimate. But in that case even the means from a balanced sample have to be weighted.

Consider the following artificial example:[1] 5 varieties A ... E of potatoes were compared. A and B were two different sorts of the same red variety, C and D are in the same way sorted from a white variety. E is the unsorted white variety hardened against frost. The same amount (in kilos) of each variety was used in the experiment. What contrasts should be used? One is certainly interested in comparing the two qualities within variety, that is:

$$\psi_1 = \alpha_A - \alpha_B$$

$$\psi_2 = \alpha_C - \alpha_D$$

Then we want to compare the white and red varieties. The qualities should then be weighted with their respective frequencies

$$\psi_3 = \frac{R_A\alpha_A + R_B\alpha_B}{R_A + R_B} - \frac{R_C\alpha_C + R_D\alpha_D}{R_C + R_D}$$

These frequencies can easily be estimated from the sorting of the potatoes for setting and the following estimate used

$$\tilde{\psi}_3 = \frac{\hat{R}_A y_{A.} + \hat{R}_B y_{B.}}{\hat{R}_A + \hat{R}_B} - \frac{\hat{R}_C y_{C.} + \hat{R}_D y_{D.}}{\hat{R}_C + \hat{R}_D}$$

However, the derivations of $E(\tilde{\psi}_3)$ and $D^2(\tilde{\psi}_3)$ will be complicated, as both the numerator and the denominator are estimated quantities. If all the potatoes had been set out; the average results for the white and red varieties could have been compared without using any weighting factors.

The fourth contrast of interest is

$$\psi_4 = \frac{R_C\alpha_C + R_D\alpha_D}{R_C + R_D} - \alpha_E,$$

which is not orthogonal to ψ_2 and ψ_3.

Dunn (9) has treated an illustrative example: some of the patients entering the Los Angeles County General hospital during the years 1959–61 and judged medically eligible for being cared for at home, had been classified into three groups, according to their home situation.

Quoting Dunn: "The groups were:

[1] Used in an examination problem for 2 "betyg" in Uppsala, December 1964.

Group 1—Eligible: Patients able to be cared for at home.

Group 2—No responsible person: Patients ineligible for home care because they had no person responsible for their care.

Group 3—Responsible person unable: Patients ineligible for home care because the person responsible was unable or unwilling to care for them."

The variable of observation was the prestige level of the patients' occupations.

The following three sample contrasts are considered by Dunn:

$$\hat{\psi}_1 = y_{1.} - \frac{r_2 y_{2.} + r_3 y_{3.}}{r_2 + r_3}$$

$$\hat{\psi}_2 = \frac{r_1 y_{1.} + r_3 y_{3.}}{r_1 + r_3} - y_{2.}$$

$$\hat{\psi}_3 = y_{1.} - y_{3.}$$

Each of these can be said to compare two populations, 1) eligible contra ineligible patients, 2) patients having a responsible person contra patients lacking a responsible person, 3) responsible person able contra responsible person unable.

These three sample contrasts estimate contrasts in a population with three groups from which r_1, r_2 and r_3 are proportional samples.

4.3. Derivation of estimates

In all the situations considered in this paragraph $\mathcal{A}$ will be a fixed factor with "effects" $\alpha(I)$, $I = 1, 2 \ldots a$. Here the most important deviation from the classical assumptions is that the variances are allowed to be unequal. This means that we have to estimate the variance for the estimate of the contrast $\psi = \Sigma_I c_I \alpha(I)$ from each level of $\mathcal{A}$ separately.

Illustration 1. In the sampling case there will be no problem as the observed means for factor $\mathcal{A}$ are uncorrelated. Let us consider a simple example with two factors; $\mathcal{B}$ nested within $\mathcal{A}$. The population model is:

$$y(IJ) = \mu + \alpha(I) + \beta(IJ) + e(IJ); \quad \begin{matrix} I = 1, 2 \ldots A \\ J = 1, 2 \ldots B \end{matrix}$$

and the statistical model:

$$y_{Ij} = \mu + \alpha(I) + \sum_J b_{Ij}(IJ)\{\beta(IJ) + e(IJ)\}; \quad \begin{matrix} I = 1, 2 \ldots a = A \\ j = 1, 2 \ldots b \end{matrix},$$

if we introduce the indicator variable

$$b_{Ij}(IJ) = \begin{cases} 1 & \text{if } Ij \text{ corresponds to } IJ \\ 0 & \text{otherwise.} \end{cases}$$

With the help of the usual side conditions it is easily seen that

$$E(\hat{\psi}) = E(\sum_I c_I y_{I.}) = \psi$$

The means $y_{I.}$ are uncorrelated as

$$\text{Cov}\,(y_{I.}; y_{I'.}) = \frac{1}{b^2} E\left[\sum_{jJ} b_{Ij}(IJ)\{\beta(IJ) + e(IJ)\} \cdot \sum_{jJ} b_{I'j}(I'J)\{\beta(I'J) + e(I'J)\}\right]$$

$$= \frac{1}{b^2} E\left[\sum_{jj'JJ'} b_{Ij}(IJ)\, b_{I'j'}(I'J')\, \beta(IJ)\, \beta(I'J')\right] = 0$$

which follows from independent sampling in different primary units (levels of factor $\mathcal{A}$) and the $e(IJ)$ being assumed uncorrelated.

Thus $D^2(\hat{\psi}) = \sum_I c_I^2 D^2(y_{I.})$

We have $D^2(y_{I.}) = \frac{1}{b}\left[\sigma_{e_I}^2 + \left(1 - \frac{b}{B}\right)\sigma_{A B_I}^2\right]$

$$E\left\{\frac{1}{b(b-1)} Q_{II}\right\} = E\left\{\frac{1}{b(b-1)} \sum_j (y_{Ij} - y_{I.})^2\right\} = \frac{1}{b}\left[\sigma_{e_I}^2 + \sigma_{A B_I}^2\right]$$

and thus an unbiased estimate

$$\hat{D}^2(\hat{\psi}) = \frac{1}{b(b-1)} \sum_I c_I^2 Q_{II},$$

if B is infinite.

Illustration 2. For a completely randomized experiment with one factor $\mathcal{A}$, we write the population model

$$y(IJ) = \mu + \alpha(I) + \varepsilon(J) + \eta(IJ) + e(IJ); \quad \begin{matrix} I = 1, 2 \ldots A \\ J = 1, 2 \ldots P \end{matrix}$$

and the statistical model:

$$y_{If} = \mu + \alpha(I) + \sum_J d_{If}(J)\{\varepsilon(J) + \eta(IJ) + e(IJ)\}; \quad \begin{matrix} I = 1, 2 \ldots a = A \\ f = 1, 2 \ldots r, \end{matrix}$$

where the indicator variable $d_{If}(J)$ is defined as follows:

$$d_{If}(J) = \begin{cases} 1 & \text{if the } f\text{th replication of treatment } I \text{ is on plot } J \\ 0 & \text{otherwise.} \end{cases}$$

It can be shown that, if $\hat{\psi} = \Sigma_I c_I y_{I.}$

$$D^2(\hat{\psi}) = \frac{1}{r}\sum_I c_I^2[\sigma_{e_I}^2 + \sigma_P^2 + 2\,\mathrm{Cov}_I\{\varepsilon(J);\,\eta(IJ)\} + \sigma_{AP_I}^2]$$

$$-\frac{1}{P}\left[\sum_I c_I^2\sigma_{AP_I}^2 + \sum_{I\neq I'} c_I c_{I'}\,\mathrm{Cov}\,\{\eta(IJ);\,\eta(I'J)\}\right].$$

The following notations have been used:

$$\sigma_{e_I}^2 = \frac{1}{P}\sum_J D^2\{e(IJ)\}$$

$$\sigma_P^2 = \frac{1}{P-1}\sum_J \varepsilon^2(J)$$

$$\sigma_{AP_I}^2 = \frac{1}{P-1}\sum_J \eta^2(IJ)$$

$$\mathrm{Cov}_I\{\varepsilon(J);\,\eta(IJ)\} = \frac{1}{P-1}\sum_J \varepsilon(J)\cdot\eta(IJ)$$

$$\mathrm{Cov}\{\eta(IJ);\,\eta(I'J)\} = \frac{1}{P-1}\sum_J \eta(IJ)\cdot\eta(I'J).$$

With the definition $Q_{II} = \sum_f (y_{If} - y_{I.})^2$ it is found that

$$E\left\{\frac{1}{r(r-1)}\sum_I c_I^2 Q_{II}\right\} = \frac{1}{r}\sum_I c_I^2[\sigma_{e_I}^2 + \sigma_P^2 + 2\,\mathrm{Cov}_I\{\varepsilon(J);\,\eta(IJ)\} + \sigma_{AP_I}^2]$$

This then gives us an unbiased estimate of $D^2(\hat{\psi})$ if P is infinite.

Illustration 3. For a randomized block design with one factor and the cell-frequencies r_{ij} determined by the population (see paragraph 2.3) the corresponding calculations have been made.

Here we only mention that

$$\hat{D}^2\{\sum_I c_I y_{I..}\} = \frac{1}{b(b-1)}\sum_{II'} c_I c_{I'}\cdot Q_{II'}$$

is an unbiased estimate of $D^2\{\Sigma_I c_I y_{I..}\}$ if the number of blocks B is infinite. $Q_{II'}$ stands here for $\Sigma_j(y_{Ij.} - y_{I..})(y_{I'j.} - y_{I'..})$, which are the basic sums used in the method discussed in paragraph 3.2 based on Hotelling's T^2. In both cases these sums are used because there are different variances within and covariances between the levels of the fixed factor.

These illustrations will indicate how the variances of a contrast can be estimated.[1] As soon as the fixed factor is crossed with any other factor (including the block factor), the method from illustration 3 can be used with b = the number of levels of the other factor.

When interval estimates for the population contrasts are made or hypotheses about them tested, the problem arises of how the sample contrast is distributed and multiple comparisons made. The latter will be discussed in part III, using mostly classical models and small deviations from them.

[1] Only contrasts in the levels of one fixed factor have been considered. Of course the methods can be applied even to contrasts in combinations of factor levels.

REFERENCES

1. Abt, K.: Table of expectations of mean squares in the analysis of variance for crossed classifications. U.S. Naval Weapons Laboratory Dahlgren, Virginia (1963).
2. Bárány, E.: Personal communication.
3. Bennett, C. A. and N. L. Franklin: Statistical analysis in chemistry and the chemical industry. Wiley, New York (1954).
4. Box, G. E. P.: Some theorems on quadratic forms applied in the study of analysis of variance problems: I. Effect of inequality of variance in the one-way classification. Ann. Math. Stat. 25 (1954), 290–302.
5. — II. Effect of inequality of variance and of correlation of errors in the two-way classification. Ann. Math. Stat. 25 (1954), 484–498.
6. Cochran, W. G.: Testing a linear relation among variances. Biometrics 7 (1951), 17–32.
7. Cornfield, J.: On samples from finite populations. J. Amer. Stat. Ass. 39 (1944), 236–239.
8. Cornfield, J. and J. W. Tukey: Average values of mean squares in factorials. Ann. Math. Stat. 27 (1956), 907–949.
9. Dunn, O. J.: Multiple comparisons using rank sums. Technometrics 6 (1964), 241–252.
10. Fisher, R. A.: Statistical methods for research workers. Oliver and Boyd, Edinburgh (1925).
11. Gärdefors, D.: Morphometrics of the grasshopper *Omocestus haemorrhoydalis* (*Acrididae*) in Sweden. To be published.
12. Hammar, H.: The influence of different fat diets on the fluorescence of the eye lens and of serum in alloxan diabetic rats. Acta Ophthalmologica 43 (1965), 61–74.
13. Henderson, C. R.: Estimation of variance and covariance components. Biometrics 9 (1953), 226–252.
14. Hooke, R.: Some applications of bipolykays to the estimation of variance components and their moments. Ann. Math. Stat. 27 (1956), 80–98.
15. Imhof, J. P.: A mixed model for the complete three-way layout with two random-effects factors. Ann. Math. Stat. 31 (1960), 906–928.
16. — Testing the hypothesis of no fixed main-effects in Scheffé's mixed model. Ann. Math. Stat. 33 (1962), 1085–1095.
17. Kempthorne, O.: The design and analysis of experiments. Wiley, New York (1952).
18. Kramer, C. Y.: On the analysis of variance of a two-way classification with unequal sub-class numbers. Biometrics 11 (1955), 441–452.
19. Lantz, L. G.: Studier av frövikter, groningshastighet och antal hjärtblad hos tall (*Pinus silvestris* L.) Unpublished manuscript, Lantbrukshögskolan, Ultuna (1964).

20. Pitman, E. J. G.: Significance tests which may be applied to samples from any populations: III. The analysis of variance test. Biometrika 29 (1937), 322–335.
21. Plackett, R. L.: Models in the analysis of variance. J. Roy. Stat. Soc. B 22 (1960), 195–217.
22. Satterthwaite, F. E.: An approximate distribution of estimates of variance components. Biometrics Bull. 2 (1946), 110–114.
23. Scheffé, H.: A "mixed model" for the analysis of variance. Ann. Math. Stat. 27 (1956), 23–36.
24. — The analysis of variance. Wiley, New York (1959).
25. Schultz, E. F.: Rules of thumb for determining expectations of mean squares in analysis of variance. Biometrics 11 (1955), 123–135.
26. Seeger, P.: Olika modeller för variansanalys av ett autentiskt exempel. Unpublished lic. thesis, Uppsala (1962).
27. — Några komplikationer vid en enkel tillämpning av variansanalys. Nord. tidsskr. f. ind. stat. 8 (1963), 35–40.
28. — En randomiseringsmodell i en konkret tillämpningssituation. Nord. tidsskr. f. ind. stat. 8 (1963), 80–86.
29. — Några exempel på bestämning av toleransgränser. Nord. tidsskr. f. ind. stat. 9 (1964), 47–54.
30. — Något om estimation vid trestegssampling. Statistisk Tidskrift (1964), 286–294.
31. Sukhatme, P. V.: Sampling theory of surveys with applications. Iowa State College Press, Ames, Iowa (1954).
32. Welch, B. L.: On the z-test in randomized blocks and Latin squares. Biometrika 29 (1937), 21–52.
33. Wilk, M. B.: Linear models and randomized experiments. Unpublished Ph. D. Thesis. Iowa state College Library, Ames Iowa (1955).
34. — The randomization analysis of a generalized randomized block design. Biometrika 42 (1955), 1–10.
35. Wilk, M. B. and O. Kempthorne: Fixed, mixed and random models. J. Amer. Stat. Ass. 50 (1955), 1144–1167.
36. — Some aspects of the analysis of factorial experiments in a completely randomized design. Ann. Math. Stat. 27 (1956), 950–985.
37. — Analysis of variance: Preliminary tests, pooling and linear models. Wright Air Development Center, Technical Report 55–244, Volume II (1956).
38. Zyskind, G.: On structure, relation, Σ, and expectation of mean squares. Sankhyā 24 (1962), 115–148.

Part III

THE PROBLEM OF MULTIPLE COMPARISONS

CHAPTER 1

STATEMENT OF THE PROBLEM

Testing the null hypothesis

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_a$$

with the "over-all" F-test we are taking a risk of α of rejecting the hypothesis when it is true. The value of α can be chosen low enough to make it reasonable to believe that not all means μ_i are alike if the hypothesis is rejected. This information, however, seldom satisfies an experimenter, he wants to make individual comparisons, to find out which contrasts in the means are responsible for the significant F-value. He often has to make several significance tests in order to run this down, and then the risk of obtaining at least one falsely significant result may be much higher than the risk of committing an error of the first kind in every single test. For example, N independent tests, each at significance level α, give a risk of obtaining at least one significant result of $1-(1-\alpha)^N$ if the null hypothesis is true for every test. When an experimenter wants to follow up a significant F-test he runs into the problem of multiple comparisons.

It was only at a rather late stage that it was realized what trouble could be caused making many significance analyses. Whether this constitutes a problem or not is still under debate. Strictly speaking the first method of solving what in this paper is regarded as a problem, was the analysis of variance test introduced by R. A. Fisher 1918 (21). Before that the alternative was to make many t-tests. But, as there is a desire to follow up a significant F-test by many t-tests, the problem of mass-significance comes back. In his classical book (22) 1935 Fisher recommends ordinary t-tests after a significant F-test, but after an insignificant F-test his advice is to heighten the level of significance by dividing the original one by the number of tests to be done. For example, if we want to compare the smallest observed mean with the largest of a observed means α should be divided by the number of possible pairs. The reason is that, if the null hypothesis is true (which has not been contradicted by the F-test), any pair of treatments could have been the smallest and largest observed; we are making one of $\binom{a}{2}$ equally possible tests. The method of making all possible t-tests with

unchanged levels of significance after a significant F-test is called the LSD-method of multiple comparisons. If the levels are heightened as above after a *significant* F, we shall call this Fisher's method. The method of t-testing without a foregoing F-test is often called the multiple-t-method.

According to Newman (34) Tippett's and Student's studies in the 1920's of range-methods may be considered as the predecessors of some of the modern methods of making multiple comparisons. Their aims seem to have been to replace the F-test by a method that not only is quicker but also gives more detailed information. These methods have been developed by Newman 1939 (34) and later by Keuls (30). In 1947 Duncan (7) and in 1949 Tukey (46) published methods which to-day seem to have been abandoned as being too complicated and inexact. Duncan has, however, adapted the main idea from his 1947-paper in a more easily applicable method (10). Since 1950 there has been a very intensive activity in this field and many methods have been produced.[1]

Eklund (17) has taken up the problem of mass-significance from a wider aspect than that of following up F-tests. In a paper by Eklund and Seeger (18) four types of situations with many analyses of significance are discussed. The four types are represented by the four fields in the following table:

	Same problem	Different problems
Same investigator	Type 1	Type 2
Different investigators	Type 3	Type 4

If one investigator makes several tests within one and the same problem (type 1), or if he studies several different problems in one and the same experiment (for one and the same publication), that is, type 2, the problem of mass-significance will arise. The term "same problem" is here used in a rather wide sense; within the same problem one can study how one or more factors influence one or more dependent variables. The situations, which preoccupied Eklund when developing his method (17), were those in which several dependent variables are studied, and which belong to both type 1 and 2. Type 4, when different investigators are working with entirely different problems, can hardly be considered as a mass-significance problem. When different investigators are working with the same problem,

[1] Besides the many papers in the reference list, the contributions to Technometrics in May 1965, wholly devoted to multiple comparisons, contain interesting discussions.

type 3, a real problem has undoubtedly been posed, as one can suspect significant results to be published more often than non-significant results. It is however difficult to see how this problem could be solved by any method of analysis without knowledge of all the analyses made.

The methods for multiple comparisons can be defined as methods for studying linear functions of the parameters in an anova-situation. They are used to study how one factor of interest influences one dependent variable. This can be said to be "the same problem" in a more strict sense than above and it is represented by the hatched area of type 1.

The type-1-situation also includes, for example, simple repetitive testing using new data in each test, but that will not concern us here.

Another method for solving the mass-significance problem is sequential testing. Nobody would question that making successive F-tests on for example 1, 2, 3 etc. blocks without using special sequential methods would invalidate the test procedure. Sequential tests will not be discussed here.

In the next chapter mass-significance in connection with F-tests will be discussed and in chapter 3 some of the different error rates that have been defined. Chapter 4 contains illustrations of several methods of making multiple comparisons. In chapter 5 the author motivates his choice of methods for use in practice.

CHAPTER 2

MASS-SIGNIFICANCE IN CONNECTION WITH F-TESTS

In the following, only mass-significance problems arising from analysis of variance will be considered. We shall distinguish between two situations which, as a rule, are treated differently. In a factorial design there may be several null hypotheses for main effects and interactions to be tested. This will be discussed in 2.1. The most well known mass-significance-situation is the one where we want to compare the levels[1] of one factor (or the combinations of more than one factor). Several methods of solving this problem are mentioned in 2.2. Both these situations belong to type 1, discussed in the preceding chapter, but only the situation in 2.2 belongs to the "hatched area" of multiple comparisons.

2.1. The case of many factors

In a factorial design with, for example, three factors $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$, with the number of levels a, b and c respectively, the sum of squares for treatments can be divided into 7 independent components:

		D.f.
Main effects	A	$a-1$
	B	$b-1$
	C	$c-1$
Two-factor interactions	AB	$(a-1)(b-1)$
	AC	$(a-1)(c-1)$
	BC	$(b-1)(c-1)$
Three-factor interaction	ABC	$(a-1)(b-1)(c-1)$
Treatment effects		$abc-1$

As a rule one wants to make all the 7 F-tests. This type of mass-significance problem has not attracted as much interest as the one involving one factor. It is sometimes regarded as no problem at all, at least when the tests can be considered independent. However, Kendall (29) and Davies (5) recommend heightening the significance level in agreement with Fisher (22).

[1] Note that this is another meaning of the word "level", which differs from that in significance level.

When the tests are positively correlated it is sometimes recommended, (25) (31), to choose the significance level α' in the individual tests so that $1-(1-\alpha')^N=\alpha$. Different significance levels in different tests will give $1-\pi_{i=1}^{N}(1-\alpha_i')=\alpha$. This α-value is then an upper limit to the risk of obtaining at least one falsely significant result. Brownlee (3) discusses the problem from a practical point of view. In common with many other investigators he points out that main effects are more probable to exist than two-factor interactions, which in turn are more probable than three-factor interactions and so on. Considering these more practical aspects the intelligent investigator could discover what effects are worth the most attention. For example, if the main effects A and B are found to be non-significant, this would probably cause us to heighten the level for the interaction AB. Birnbaum (2) and Daniel (4) have discussed the problem in connection with unreplicated and fractionally replicated designs.

Many of the methods presented in paragraph 2.2 could be applied even here.

2.2. The case of many levels per factor (or factor combination)

Some of the methods to be described can be applied whenever the comparisons are: (i) independent or dependent (ii) contrasts or non-contrasts. Others are constructed for contrasts only. Some of these methods are preceded by an F-test and applied only when the F-test gives a significant result, other methods are used directly.

Some of the methods will be described and commented upon in chapters 4 and 5. At this time we shall only list them.

When a method is constructed so that the critical value for a given type of contrast (for example, contrasts in pairs of means) is the same irrespective of which means are included in the contrast, then we say that it has a constant least significant difference. Another group of methods to be treated separately is one where the critical value depends on how many of the means that have numerical values between the two means in a pairwise contrast (a contrast including a pair of treatments). A comparison of the largest and smallest means will require a larger critical value than a comparison of two means that are consecutive to each other in a rank order by size. These methods are said to have a variable least significant difference. We shall also distinguish between methods for comparing with a standard treatment and methods to be used when the classical assumptions are not satisfied. To complete the picture three more methods will be described. Two of them, Duncan's (1947) and Eklund's could be included

in the groups mentioned above, but Duncan's method does not seem to be used much (it has been replaced by Duncan's new method 1955), and Eklund's is very different from the others though it can be said to have a constant least significant difference.

A. Methods with constant least significant difference:
   1) Scheffé's S-method
   2) Tukey's T-method (from 1951)
   3) Multiple-t-method
   4) LSD-method
   5) Fisher's method

B. Methods with variable least significant difference:
   6) Newman–Keuls' method
   7) Tukey's method (from 1953)
   8) Duncan's method (from 1955)

C. Method for comparisons with a standard:
   9) Dunnett's method

D. Some non-parametric methods:
   The sign test:
   10) Steel's method for making comparisons with a standard
   11) Fisher's method for making any sign test
   t-tests with individual estimates of variances:
   12) Fisher's method
   Rank sum tests:
   13) Steel's method for making comparisons with a standard
   14) Steel's and
   15) Dunn's methods for making more general comparisons

E. Two older methods:
   16) Tukey's method (from 1949)
   17) Duncan's method (from 1947, 1951)

F. 18) Eklund's method.

CHAPTER 3

DIFFERENT OPERATIONAL DEFINITIONS OF TYPE-I-ERROR FREQUENCIES

Before we start describing the various methods of making multiple comparisons it would be useful to study some of the different specifications of risks of committing a type-I-error. With a type-I-error we mean that either a true null hypothesis has been rejected or that a false null hypothesis has been rejected in the wrong direction. Four specifications, A–D, are given below:

A. At most a certain proportion α of all tests, in the universe of all possible tests, is allowed to have false significances. This is denoted in (26) by "error rate per comparison". Such a definition is used in the multiple-t-method and the LSD-method, that is when the t-tests are used without heightening the level of significance. Thus for every test with true null hypothesis the probability of committing an error of the first kind is α. If, however, we make several tests with true null hypotheses the probability of at least one error may be much larger than α.

B. At most a certain proportion of all experiments (including several tests) is allowed to contain one or more false significances (familywise or experimentwise error rate). For the separate experiment no specification is made. Here the number of tests per experiment is important. The greater this number is, the greater will be the difference between error rate per comparison and experimentwise error rate. This definition is used for example with the S- and T-methods.

C. At most a certain (expected) number of all tests in an experiment may have false significances (error rate per family). If, as with Fisher's method, one divides the total level α by the number of tests N to get the level in the individual test, one is using error rate per family.

Let us introduce the variables x_i; $(i = 1, 2 \ldots N)$ with the following definition:

$$x_i = \begin{cases} 1 & \text{when test no. } i \text{ gives a false significance} \\ 0 & \text{otherwise} \end{cases}$$

Then $X = \sum_{i=1}^{N} x_i$ denotes the number of false significances in the experiment and has the mean value

$$E(X) = N \cdot \frac{\alpha}{N} = \alpha$$

if the null hypothesis is true for all of the N tests of significance. If the null hypothesis is true for only N_0 tests we have

$$E(X) \leqslant N_0 \cdot \frac{\alpha}{N} + (N - N_0)\frac{\alpha}{2N} = \alpha \cdot \frac{N + N_0}{2N}; \quad N_0 \leqslant N$$

Scheffé (37) and Dunn (13) have pointed out, that according to Bonferroni's inequality

P (no significant result $| H_0$ is true in every test) $\geqslant 1 - N \cdot \alpha/N$

or P (at least one significant result $| H_0$ is true in every test) $\leqslant \alpha$.

Thus we have an upper limit for the risk level of type B. According to C Fisher's method has the level of significance α, but according to B α is an upper limit.

D. At most a certain proportion of all significances in the experiment may be false. If for example n significant results are observed we shall know that at most $k \cdot n$ are false (on the average) where k is some small factor such as 0.01. This requirement was introduced by Eklund (17).

Newman–Keuls' (34, 30) and Duncan's (10) methods are so different that they are not discussed here. It is difficult to characterize their type-I-risks as easily as has been done in A–D.

Eklund's method is based on an entirely different operational definition of type-I-errors than the other methods and has therefore different characteristics. For example, if the number of true significances is in some way increased, more false significances can be tolerated.

It has been much discussed in the literature on this subject whether error rate per comparison or experimentwise error rate should be used. I.e. whether or not one should bother about the mass-significance problem, and in which situations. See for example, Steel (43), Harter (26) and Gabriel (23).

Let us consider a type of experiment, namely consumer testing, which will be much used below to exemplify different methods. Testing a common article, for example washing-machines, watches or cars, rather large investigations have to be made in order to cover the market at least partly. Even if the investigation is going on stepwise, with a small number of objects tested at the same time, it must be so arranged that all of them

can be compared. The mass-significance problem will thus continue to exist. One common practice to present the results obtained from the testing of a certain characteristic is to divide the ranked means, more or less subjectively, into classes of the type "very good", "not so good" and so on. But if the differences between, for example, all the means in the classes "very good" and "good" can be explained by random errors the classification will be misleading. Thus arises the need for analyses of significance. Sometimes it is argued that a potential buyer choosing between two brands does not bother about how many other brands have been investigated. And it is obvious that if for some reason, say the price, the choice lies between only two brands, the problem of mass-significance will not arise. But the person who is going to establish a ranking order as correctly as possible will have the problem of mass-significance. If he does not consider that many analyses have been made or could be made, he may reach an unjust verdict. How the results of consumer testing should be presented so as to satisfy both of these requirements (in a way the consumer's and the producer's requirements) is not an easy problem. Perhaps the consumer will have to be taught enough statistics to be able to judge for himself.

CHAPTER 4

REVIEWS AND ILLUSTRATIONS OF THE EXISTING METHODS

In this chapter short descriptions of the methods mentioned in chapter 2.2 for making multiple comparisons are given. The aim is to make it possible for the reader to use them, at least in uncomplicated situations, and to give a background for the practical discussion in the next chapter. Detailed comparisons of some of the methods can be found in papers by Duncan (10), Gabriel (23) and Harter (26). References to the theoretical derivations are also given throughout the chapter.

4.1. Methods with constant least significant difference

A numerical example.—The cleaning ability of 9 types of washing-machines was compared in 15 different trials (blocks) in order to advise consumers (38). For every combination of machine and trial the whiteness of three objects (for example pillowcases) was measured.

The analysis can be summarized as follows:

Analysis of variance table

Source of variation	D.f.	Mean square	F	$F_{.0005}$
Trials	14	65.39	—	—
Machines	8	55.88	26.23[1]	~3.8
Interaction	112	2.13	1.97	~1.9
Residual	270	1.08	—	—

Mean value table

Wh	B	V	C	E	Z	R	WII	WI
45.4	46.9	47.8	48.0	48.0	48.1	48.2	49.0	49.0

[1] The model is mixed with "machines" fixed and "trials" random.

a) *The S-method*

Scheffé's S-method (36, 37) is built on the F-distribution and constructed in such a way that at least one linear contrast will be significantly different from zero if and only if the over-all F-test gives a significant result. However, sometimes it is difficult to find the significant contrasts. A (numerically laborious) method to find all of them is given in Gabriel (23).

Let us say that the contrast in the population means

$$\psi = \sum_{i=1}^{a} c_i \mu_i; \quad \sum_{i=1}^{a} c_i = 0$$

is estimated by the sample contrast

$$\hat{\psi} = \sum_{i=1}^{a} c_i \bar{y}_i$$

where c_i are known constants and $\bar{y}_i$ observed treatment means. If the $\bar{y}_i$ are independently and normally distributed with variances $k_i\sigma^2$ then the probability is $1-\alpha$ that the following criterion is simultaneously satisfied for every possible contrast

$$|\hat{\psi} - \psi| \leqslant \hat{\sigma} \sqrt{(a-1)\, F_{\alpha}^{a-1;\nu} \cdot \sum_{i=1}^{a} c_i^2 k_i}$$

Here ν is the number of degrees of freedom for the estimate $\hat{\sigma}^2$ of σ^2. The k_i's are supposed to be known; if $\bar{y}_i$ is a mean of r_i independent observations and all $\Sigma_i r_i$ observations in the sample have the same variance σ^2 then $k_i = 1/r_i$.[1]

To compare the mean of WI and WII $[(\mu_{WI}+\mu_{WII})/2]$ with the mean of the 7 other machines at a level of 5% one computes:

$$\hat{\psi} = \frac{49.0+49.0}{2} - \frac{48.2+\ldots+45.4}{7} = 1.5$$

and obtains the least significant difference for this contrast,

$$\hat{D} = \hat{\sigma} \sqrt{(a-1)\, F_{.05}^{8;112} \cdot \frac{1}{r} \sum_i c_i^2} = \sqrt{2.13} \sqrt{8 \cdot 2.02 \cdot \frac{1}{45}\left(\frac{1}{2}+\frac{1}{7}\right)} = 0.7$$

At this level we can thus say that

$$\psi > 0$$

[1] The S-method is in fact even more general in that it is applicable to all linear combinations of the parameters in the experiment.

In fact with at least 95%-s confidence

$$0.8 \leqslant \psi \leqslant 2.2.$$

All possible comparisons in pairs would give the least significant difference:

$$\hat{D}_S = \sqrt{2.13}\sqrt{8 \cdot 2.02 \cdot \tfrac{2}{45}} = 1.24$$

and thus

WI, WII, R > B, Wh

Z, E, C, V, B > Wh

where > means "significantly greater than".

In the following, least significant differences for pairwise contrasts will be computed by several methods. For the S-method it is denoted by $\hat{D}_S$, for the T-method by $\hat{D}_T$ etc.

For pairwise comparisons, however, the S-method has lower power than the T-method.

b) *The T-method*

This method, originated by Tukey (47), utilizes the distribution of the Studentized range, that is, the range in the sample divided by the standard deviation computed from the sample. The criterion of the T-method is:

With probability $1-\alpha$ all possible contrasts simultaneously satisfy:

$$|\hat{\psi} - \psi| \leqslant q_{\alpha;\, a;\, \nu} \cdot \hat{\sigma}_{\bar{y}} \left(\tfrac{1}{2} \sum_{i=1}^{a} |c_i| \right)$$

where $q_{\alpha;\, a;\, \nu}$ is the upper α-point in the distribution of the Studentized range[1] and $\hat{\sigma}_{\bar{y}}$ the estimated standard deviation of the treatment means.

This means that the standard deviations of the means have to be equal; the assumptions are otherwise the same as in Scheffé's method. Tukey's method is primarily intended for pairwise comparisons, for which:

$$\tfrac{1}{2} \sum_{i=1}^{a} |c_i| = 1$$

The T-method applied to pairwise comparisons in the example above gives

$$\hat{D}_T = q_{.05;\, 9;\, 112} \cdot \hat{\sigma}_{\bar{y}} = 4.48 \cdot \sqrt{\frac{2.13}{45}} = 0.98$$

and

[1] Tabulated, for example, in (35).

WI, WII > C, E, V, B, Wh

R, Z, C, E > B, Wh

V, B > Wh

The T-method is more powerful than the S-method using pairwise comparisons, but the latter is more robust. Comparisons between these two methods can be found in (36) and (37).

c) *The multiple-t- and LSD-methods*

As has been said before, the difference between these methods is that the multiple-t-method consists in making an arbitrary number of ordinary t-tests instead of an F-test, while in the LSD-method the t-tests are only made after a significant F-test. Probably all investigators using statistical methods have long ago replaced the multiple-t-method by the LSD-method. The LSD-method has been much used, but is now gradually being abandoned as it is considered to give too many false significances. The name LSD comes from the fact that a least significant difference for a certain type of contrast is computed and compared with the observed values of the contrasts. This was mostly done for pairwise comparisons of means with equal variances, in which case the LSD is $t''_\alpha \cdot \hat{\sigma}_{\bar{y}} \cdot \sqrt{2}$. When the means have unequal variances, the LSD varies even between contrasts of the same type and it requires a little more effort to find the significant contrasts.

In principle the S- and T-methods apply the same procedure as the LSD-method, but the least significant differences are computed in different ways.

Another similarity between the T- and LSD-methods is that they both utilize the Studentized range. For example, the LSD-method applied to pairwise comparisons gives:[1]

$$\hat{D}_{\text{LSD}} = t''_\alpha \cdot \hat{\sigma}_{\bar{y}} \cdot \sqrt{2} = q_{\alpha;\,2;\,\nu} \cdot \hat{\sigma}_{\bar{y}}$$

which is to be compared with:

$$\hat{D}_T = q_{\alpha;\,a;\,\nu} \cdot \hat{\sigma}_{\bar{y}}$$

for the T-method.

These methods can be considered as two variants of the same thing. The LSD-method always involves the risk α of considering two population means different when they are not, but the risk of making at least one false significant conclusion increases with a, the total number of means con-

[1] For two means $\bar{y}_1$ and $\bar{y}_2$ both on r observations $t = (\bar{y}_1 - \bar{y}_2)/\sqrt{2\hat{\sigma}^2_{\bar{y}}}$ and $q = (\bar{y}_1 - \bar{y}_2)/\hat{\sigma}$ - = range/standard deviation.

sidered. The T-method always involves the risk α of reaching at least one false significance, but the risk for individual pairs decreases with the increasing a. Of course we could make the methods exactly alike by choosing different α-values for them. It is not obvious that we should use the same α with error rate per comparison (LSD)[1] as with experimentwise error rate (T).

For the numerical example with $\alpha = 0.05$ we obtain

$$\hat{D}_{\text{LSD}} = t'_\alpha \sqrt{2\hat{\sigma}_{\bar{y}}^2} = 1.98 \sqrt{\frac{2 \cdot 2.13}{45}} = 0.61$$

and the significant differences

$$\text{WI, WII} > \text{R, Z, E, C, V} > \text{B} > \text{Wh.}$$

Here we get a proper division of treatments, just the type of detailed information we want, and then of course it is tempting to use the LSD-method. This method is probably still in use, but it is no longer recommended in text books because of its large experimentwise error rate.

d) *Fisher's method*

This method differs from the other ones reviewed here, in that it only specifies the individual level of significance, not the test variable. It can be applied to t-tests, sign tests, rank tests etc. Eklund's method (see paragraph 4.6) is in this respect of the same type as Fisher's method.

If m comparisons are planned before the investigation, they are tested, using this method, at level α/m. The total level of significance is then at most α (experimentwise). This method was suggested by Fisher (22) and has since then been recommended in text books, for example Kendall (29) and Scheffé (37). Dunn (13) has discussed it thoroughly and compared its power with that of the S-method. She found that Fisher's method was, for the most part, more sensitive, but for some values of m and a less sensitive than Scheffé's. In a recent paper (39) Siotani discusses the possibility of constructing intervals with a confidence level of exactly $1-\alpha$ instead of $>1-\alpha$, as is usually the case using Fisher's method. If this is possible, Fisher's method will never give greater LSD's than the S-method. Siotani gives approximate tables for some special cases.

Let us say that in the investigation of the washing-machines we have planned to make all pairwise comparisons, that is $\binom{9}{2} = 36$ in number. The

[1] The LSD-method does not exactly use the risk α as error rate per comparison, because it starts with a significant F-test.

level to be used in the individual tests is thus $\alpha' = \alpha/36$. If we want to find the critical t-value easily, it is feasible to choose $\alpha' = 0.001$ or 0.002. This means that the total significance level is at most equal to $\alpha = 0.036$ or 0.072. Some of the tests (for example of $\bar{x}_1 - \bar{x}_2$ and of $\bar{x}_1 - \bar{x}_3$) are positively correlated and thus the experimentwise error rate is certain to be less than α. Therefore justification can be found for choosing the larger of the two values. With $\alpha' = 0.002$ we get:

$$\hat{D}_F = t'_{\alpha'} \sqrt{2 \cdot \hat{\sigma}^2_{\bar{y}}} = 3.17 \sqrt{\frac{2 \cdot 2.13}{45}} = 0.98$$

The same result is obtained here as with the T-method above.

4.2. Methods with variable least significant difference

The three multiple range tests to be illustrated in this paragraph are Newman–Keuls' (34, 30), one method originated by Tukey 1953 (48) and Duncan's new multiple range test (10). Like the T-, LSD- and multiple-t-methods discussed in the preceding paragraph they all utilize the distribution of the Studentized range, but the critical values of the methods discussed in this paragraph depend on the difference in rank for the treatments, which have been ranked after the magnitude of their means. These 6 range tests have been illustrated by Harter (26), who also compares their respective risks for type-I- and type-II-errors.

a) *Newman–Keuls' method*

The observed treatment means are arranged in order of magnitude and we start by comparing the largest with the smallest; that is, the significance for the range of a means is investigated.[1] For significance it is required that

$$|\bar{y}_{\max} - \bar{y}_{\min}| > q_{\alpha;\, a;\, \nu} \cdot \hat{\sigma}_{\bar{y}}$$

If this criterion is satisfied the comparison continues; the largest mean is compared with the next smallest and the next largest with the smallest; the significance for the range of $a-1$ means is thus investigated. If at least one of these ranges is found significant, ranges of $a-2$ means are being tested and so on until no more significances are obtained. No range included in an insignificant range is accepted as significant. The critical value for testing the range of p means is $q_{\alpha;\, p;\, \nu} \cdot \hat{\sigma}_{\bar{y}}$. In the preceding paragraph we saw that with the T-method the critical value is always $q_{\alpha;\, a;\, \nu} \cdot \hat{\sigma}_{\bar{y}}$, and with the LSD-method always $q_{\alpha;\, 2;\, \nu} \cdot \hat{\sigma}_{\bar{y}}$.

[1] This common phrase means that we test the hypothesis H_0: $\mu_1 = \mu_2 = \ldots = \mu_a$ (i.e., that the population range is zero).

b) *Tukey's method from 1953*

This method is identical with Newman–Keuls', except that the critical value for testing a range of p means is $\frac{1}{2}(q_{\alpha;\,p,\nu} + q_{\alpha,\,a,\,\nu}) \cdot \partial_{\bar{y}}$, the arithmetic mean of the critical values for Newman–Keuls' method and the T-method.

c) *Duncan's new multiple range test*

Duncan's method differs from the other two herein discussed by the choice of critical value, in that it uses:

$$R_{\alpha;\,p;\,\nu} \cdot \partial_{\bar{y}}$$

where: $$R_{\alpha;\,p;\,\nu} = \max \left\{ q_{1-(1-\alpha)^{(p-1)};\,p;\,\nu},\, q_{1-(1-\alpha)^{(p-2)};\,p-1;\,\nu} \right\}$$

Tables of $R_{\alpha;\,p;\,\nu}$ have been published by Duncan (10), and corrected and extended by Harter (27, 28).

The motivation for the choice of $R_{\alpha;\,p;\,\nu}$ seems to be as follows: a comparison between p means (at least with infinitely many degrees of freedom for the estimate of the variance[1]) can be divided into $p-1$ independent comparisons. Duncan means that with $p-1$ independent tests "no reasonable objection could be raised to the joint unmodified application of $p-1$ 5 % level tests". The probability of at least one significant result when all population means are equal is then $1-(1-\alpha)^{p-1}$. As this is reasonable it would also be reasonable to use the significance level $1-(1-\alpha)^{p-1}$ (protection level $(1-\alpha)^{p-1}$ in Duncan's terminology) in testing the range of p means. This method of choosing the level of significance will give, for increasing a, Duncan's method a much higher power than the other methods. On the other hand the more identical means we include, the higher will be the probability $1-(1-\alpha)^{a-1}$ of at least one false significance. In practice, however, it is probable that the more treatments we include in the experiment the more real differences we get.

Numerical results

The following critical values are obtained from the investigation of the washing-machines:

Method	Number, p, of treatments in range							
	2	3	4	5	6	7	8	9
LSD	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61
Duncan	0.61	0.64	0.66	0.68	0.69	0.70	0.71	0.72
Newman–Keuls	0.61	0.73	0.80	0.85	0.89	0.92	0.95	0.98
Tukey's 1953	0.79	0.85	0.89	0.91	0.94	0.95	0.96	0.98
T	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
Fisher	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
S	1.24	1.24	1.24	1.24	1.24	1.24	1.24	1.24

For Fisher's method $\alpha = 0.072$ for the others $\alpha = 0.05$.

[1] Finney (20) has shown that 20 d.f. give a fairly good approximation.

The resulting significant differences not given before are, from Newman–Keuls and Duncan:

$$\text{WI, WII} > \text{R, Z, E, C, V} > \text{B, Wh}$$

from Tukey:

$$\text{WI, WII} > \quad \text{Z, E, C, V, B, Wh,}$$

$$\text{R, Z, E, C, V} > \text{B, Wh}$$

We see that Newman–Keuls' and Duncan's methods give the same result as the LSD-method.

An approximate method to extend these three tests, and the T-method, to the case where the means have different numbers of observations is mentioned by Kramer (32). If $\bar{y}_1$ and $\bar{y}_2$ are computed from r_1 and r_2 observations respectively, he substitutes $\frac{1}{2}(1/r_1+1/r_2)\hat{\sigma}^2$ for $\hat{\sigma}^2_{\bar{y}}$ when $\bar{y}_1-\bar{y}_2$ is studied and so on. According to Kramer this makes the tests more conservative. In a later paper (33) Kramer generalized this to correlated means. Duncan (11) "presents a more complete method for these extensions which necessarily sacrifices a little in simplicity but is more powerful, especially in cases in which the differences between the means have appreciably different variances." This more complete method consists of the following: If the variances for the differences between two means are written $D^2(\bar{x}_i-\bar{x}_j)=k_{ij}\cdot\sigma^2$, adjusted differences $(\bar{x}_i-\bar{x}_j)\cdot\sqrt{2/k_{ij}}$ are introduced. The rule is then: "Any subset of p means is homogeneous if the largest adjusted difference in the subset fails to exceed the critical value." Kramer's method considers unadjusted ranges and may thus give fewer significant results.

For the S- and LSD-methods as well as for Fisher's method no special corrections for unequal numbers of observations are necessary.

4.3. Method for comparisons with a standard

Suppose that an investigation includes p experimental treatments and one standard treatment. The main interest may be to make all p comparisons with the standard. Dunnett (15, 16) has published a method specially constructed for this purpose. It is based on the positive correlation between the p tests; positive because they all include the standard. Dunnett's tables are computed for the correlation $+0.5$ (when there is an equal number of observations for all the treatments including the standard), but a correction method is given in the later paper (16), to be applied when more observations are made of the standard treatment than of the other treatments.

TABLE III: 1. *Primary data and analysis of variance table for a psycho-physical experiment (24).*

Individual no.	Tone A	B	C	D
1	220.3	220.5	219.7	220.5
	220.3	220.3	219.7	220.5
2	220.7	220.7	220.3	220.3
	220.0	220.3	220.3	220.3
3	220.0	219.7	220.5	220.3
	220.0	220.0	219.7	219.7
4	219.7	220.7	220.3	220.5
	220.5	220.3	220.0	221.7
5	219.7	219.5	219.5	219.3
	220.0	219.5	220.3	219.7
6	220.3	221.5	220.5	221.5
	220.0	221.0	220.5	221.7
7	220.0	219.3	219.5	220.5
	220.0	219.7	220.3	220.0
8	220.0	220.5	219.7	220.5
	219.7	221.5	220.0	220.7
9	218.7	220.5	221.0	221.0
	220.5	220.7	219.7	221.7
10	220.0	220.0	219.7	220.7
	220.0	220.5	220.0	221.0
Mean	220.02	220.34	220.06	220.61

Analysis of variance table:

Sources of variation	D.f.	Mean squares
Individuals	9	0.89
Tones	3	1.48
Interaction	27	0.28
Residual	40	0.17

A numerical example.—In a psycho-physical experiment (24), the individuals tested were asked to adjust the frequency of four different tones to that of a standard tone. One of the four tones was the standard itself. The problem was to find out whether the comparison A, the standard with itself, differed from the three other comparisons, denoted by B, C and D. Ten individuals made two adjustments on each of the four tones. The results are shown in table III: 1.

Making the three double-sided tests at the 5%-level we obtain, with the help of Dunnett's table (16), an approximate[1] least significant difference:

$$2.49\sqrt{\frac{2 \cdot 0.28}{20}} = 0.42$$

The only significant difference is thus:

$$D > A.$$

4.4. SOME NON-PARAMETRIC METHODS

The methods discussed thus far assume, at least approximately, normal distributions and homogeneity of variances. When the deviations from these ideal conditions become too great the tests may work badly. Therefore

[1] Linear interpolation in Dunnett's table.

some non-parametric methods have been adjusted for use in making multiple comparisons. Kramer's (33) and Duncan's (11) methods with adjusted differences also assume a sort of variance homogeneity. They assume known correlations and known ratios of variances and therefore the methods do not belong to this paragraph. The *t*-tests with individual estimates of variances, however, are used when there is a lack of variance homogeneity, and they will be treated here.

a) *The sign test*

Steel (42) has studied the sign test as a method of making all possible comparisons with a standard in randomized block designs. He computes some exact sampling distributions and suggests that Dunnett's tables modified for the sign test, can probably be used for the analysis of significance.[1]

Let us apply this method to the numerical example in paragraph 4.3. There we had 3 experimental treatments and 10 blocks. From Steel's double-sided table at the level 5% (42) we obtain the critical value 0; that is only if all the signs are equal will the difference be said to be significantly different from zero. In such cases it is often very easy to reach the verdict "not significant". We get:

A–B	*A–C*	*A–D*
–	+	–
–	+	+
+	–	
not sign.	not sign.	not sign.

if we write down the signs of the average differences for the individuals.

Thus no significant result was found above; in fact much broader experiments must be used to get a satisfactory sensitivity.

Of course it is possible to use the sign test for more general comparisons, for example all pairwise comparisons. Fisher's method of dividing the total level of significance with the number of tests is then utilized. To be able to illustrate this by using a somewhat broader experiment as a basis, we go back to the investigation of 9 washing-machines in 15 blocks. In that experiment we wanted to make the 36 tests of all possible pairs of treatments, and if we choose the critical value 0 as above, an upper limit for

[1] In a paper in Technometrics (August 1965) Phyne and Steel have published somewhat revised and extended tables.

the total significance level α will be: $36 \cdot 2 \cdot (\frac{1}{2})^{15} = 0.0022$.[1] The critical value 1, will in the same way give: $\alpha \leqslant 0.0353$, and the critical value 2: $\alpha \leqslant 0.27$.

It is then reasonable to allow one different sign in a significant comparison, which gives:

WI > V, C, B, Wh

WII > C, B, Wh

R, Z, E, V, C > Wh

a result to be compared with those obtained in paragraphs 4.1 and 4.2. Note that we here arrive at one more significant difference than the *S*-method gave.

One problem with the sign test is the possible appearance of ties; that is, when we observe neither + nor − but 0. One common method is to set a zero as equal to half a + and half a −, another method is to ignore the ties and reckon with a reduced number of blocks.

Using the sign test we need assume neither normality nor equality of variances, and the test works even when the distributions differ from block to block. But as in most cases with non-parametric tests the existence of interactions may cause trouble. (See part IV, chapter 3.3.1, where the possibility of using the sign test on more complicated contrasts is also discussed.) However, we must use such a great number of blocks in order to get the sign test satisfactorily sensitive, that the mean value of the differences may be regarded as approximately normally distributed. Besides, if the distributions for the differences are the same in every block, which they ought to be if the blocks are of about the same type, it is possible to use:

b) *t-tests with individual estimates of variances*

With individual estimates we mean that the variance is estimated only by considering the treatments in the contrast. Using randomized blocks the value of the contrast is computed for every block, and thereafter the variance between these values. This procedure is more sensitive than the one using the sign test, and it tests quantitative null hypotheses while the sign test tests a qualitative null hypothesis. (See part IV.) The *t*-test like the sign test will always have interaction as an error term, which means, that if the model is fixed, the significance will be underestimated. These

[1] This value is obtained with the help of the binomial distribution and the hypothesis of equal probability for plus- and minus-signs.

t-tests are applied to pairwise comparisons in the investigation of the washing-machines with the following significant differences:

$$\begin{aligned} \text{WI} &> \text{E, V, C, B, Wh} \\ \text{WII} &> \text{C, B, Wh} \\ \text{R} &> \text{B, Wh} \\ \text{Z, E, V, C, B} &> \text{Wh} \end{aligned}$$

The individual levels were $\alpha' = 0.001$, with the total level $\alpha \leqslant 0.036$, in order to make it comparable with the sign test, where $\alpha \leqslant 0.035$.

c) *Rank tests*

Steel (41) has also considered completely randomized designs with one factor, and in the same way as for the sign test made it plausible that Dunnett's tables, suitably modified, can be used for the rank test in making comparisons with a standard. In this test one ranks only the observations made on the pair of treatments in the comparison. Equally many observations per treatment are required. The smallest of the rank sums for the two treatments is compared with a critical value that can be taken from Steel's tables or, for small groups, computed.

Steel (44) has extended the above rank test so that it can be used for comparing all possible pairs of treatments. Exact sampling-distributions are given for three treatments and up to six observations per treatment.

A more general case has been treated by Dunn (14) who considers arbitrary contrasts

$$y_m = \frac{\sum_i T_i}{\sum_i r_i} - \frac{\sum_{i'} T_{i'}}{\sum_{i'} r_{i'}}; \quad m = 1, 2 \ldots p$$

where T_i is the rank sum and r_i the number of observations for treatment no. i. The ranking can be made either for all the observations together or for the treatments in the contrast only (Steel's procedure). Thus the rank sums in the last procedure may vary from contrast to contrast, while they are constant in the first. In both cases the variances for the contrasts can be written as:

$$\sigma_m^2 = \left\{ \frac{R(R+1)}{12} - \frac{\sum_{s=1}^{v} (t_s^3 - t_s)}{12(R-1)} \right\} \cdot \left\{ \frac{1}{\sum_i r_i} + \frac{1}{\sum_{i'} r_{i'}} \right\}$$

Here R is the total number of observations, or the number of observations in the contrast, depending on which ranking system is applied. Now t_s is the number of ties in the sth group of v groups of ties. If the number of

observations is large, y_m/σ_m may be compared with the critical values $\pm\lambda_{\alpha/p}$, which means that Fisher's method has been used. Note that the two procedures use the same formulas but that the T_i, R and t_s may be different. The difference between the two procedures seems to be slight (14).

The rank sum test assumes identical distributions. Thus the only "normal-theory-assumption" avoided is that of normal distribution, an assumption which loses in importance when the number of observations increases. Therefore when the variables are quantitative, the gain in robustness by using rank tests instead of t-tests is even smaller than by using the sign test. However, when the data consists only of ranks, this often leaves us with the rank test as the sole alternative.

4.5. Two older methods

For the sake of completeness two additional methods will be described, one by Tukey 1949 (46) and Duncan's multiple comparisons test from 1947 (7) and 1951 (8).

a) *Tukey's method*

Tukey's method is based on three types of tests and is thus rather complicated. The first step is to arrange the means in their order of size and then test for excessive gaps between adjacent means using ordinary t-tests. In this way the treatments are divided into separate groups. Repeated tests for stragglers (outliers) are then applied to groups with at least three means. Stragglers found significantly different from the group means are broken off from the group. Finally F-tests for excessive spread are applied to the resulting subgroups. Tukey (46) makes the conjecture that, "if the same level, such as 5% is used in all three tests, the frequency of false positives will be between 1.2 and 1.6 times the level used (i.e., between 6% and 8% when a 5% level is used)", a conjecture that seems a bit optimistic.

The method is illustrated by applying it to the washing-machine experiment used above. The first step, as in the LSD-method, will result in the following partitioning into groups:

$$\text{WI, WII} > \text{R, Z, E, C, V, } > \text{B} > \text{Wh}$$

For the group R, Z, E, C, V we then make the straggler test in which the test variable is:

$$\frac{\dfrac{\max|\bar{y}_i - \bar{\bar{y}}|}{\hat{\sigma}_{\bar{y}}} - \dfrac{6}{5}\log_{10} a}{3\left(\dfrac{1}{4} + \dfrac{1}{\nu}\right)} \qquad \text{when the number of means } a > 3$$

or

$$\frac{\dfrac{\max|\bar{y}_i - \bar{\bar{y}}|}{\hat{\sigma}_{\bar{y}}} - \frac{1}{2}}{3\left(\frac{1}{4} + \frac{1}{\nu}\right)} \quad \text{when } a = 3$$

where $\bar{\bar{y}}$ is the group mean and ν the number of degrees of freedom for $\hat{\sigma}_{\bar{y}}$. The test variables are considered distributed approximately as $N(0, 1)$. When a significant result is obtained the straggler is broken off from the group, and the test is repeated on the rest of the group as long as it contains at least three means. Stragglers broken off in the same direction constitute a new group, and if this group contains at least three means, it is tested for stragglers and so on. In the numerical example we obtain:

$$\frac{\dfrac{|47.8 - 48.02|}{\sqrt{\dfrac{2.13}{45}}} - \dfrac{6}{5} \cdot \log 5}{3\left(\dfrac{1}{4} + \dfrac{1}{112}\right)} < 1$$

Thus the second step gives no further significance, nor does the F-test in the third step. Tukey's method and the LSD-method (and Newman–Keuls' and Duncan's range tests) thus give the same result for this example. This three-stage-method has probably not been much used; the only textbook known to the author in which it has been described is Adams' (1).

b) *Duncan's method*

Duncan's multiple comparisons method utilizes, in principle, the same ideas as his new multiple range test described above, but it uses F-tests instead of range-tests. The levels of significance are the same: $\alpha_p = 1 - (1 - \alpha)^{p-1}$; $p = 2, 3, \ldots a$. In a restricted form of the test, when we want to find out which pairs of treatments are significantly different, the F-tests are applied in exactly the same way as in the multiple range test. Tables with $\alpha = 0.05$ and 0.01 can be found in (8). A two-stage routine has been developed in order to make approximate but more rapid calculations. The complete form of the test compares not only pairs but even groups of means, and thus possesses a somewhat higher power than the range test. This means that a third stage is included in the routine. This method is rather complicated and time-consuming. I believe most investigators prefer the range test.

4.6. A discussion of Eklund's method

As has been mentioned in chapter 3, Eklund's method has been designed to reduce the ratio of false significances to the total number of significances.

Consider the variable:

$$y = \frac{\text{number of false significances}}{\text{number of significances}} = \frac{u}{v}$$

where v is observed but u has to be estimated with the help of the individual significance level α', and the largest number of tests in which the null hypothesis may be true. The problem is then to determine α' so that $y \leqslant k$, where k is a predetermined level.

Let us assume that we want to make N analyses of significance, each analysis at the level of α', and that the null hypothesis is true for $N_0 \leqslant N$ of them. When the null hypothesis is true the probability of a false significance $= \alpha'$, otherwise it is $< \alpha'/2$ (double-sided test). Thus we have

$$E(u) \leqslant N_0 \cdot \alpha' + (N - N_0) \cdot \frac{\alpha'}{2} = \alpha' \cdot \frac{N + N_0}{2} \quad \text{(equality when } N_0 = N\text{)}$$

The right part of this inequality is taken as a sort of estimate for u. The total number of significances at the individual level α' is denoted by $n(\alpha') = N \cdot p(\alpha')$. Then the following criterion is obtained:

$$y \approx \frac{\alpha' \cdot (N + N_0)/2}{N \cdot p(\alpha')} \leqslant k$$

N_0 is of course unknown and one can either reckon with the most unfavourable case $N_0 = N$, which gives a conservative α', or estimate N_0 from the obtained results, for example, setting

$$\hat{N}_0 = N\{1 - p(\alpha') + \alpha'\}$$

The first alternative gives the quotient:

$$\frac{\alpha'}{p(\alpha')} \leqslant k$$

the second:

$$\frac{\alpha'\{2 - p(\alpha') + \alpha'\}}{2p(\alpha')} \leqslant k$$ [1]

[1] Eklund (17) discusses the conventions

$$\frac{\alpha'}{p(\alpha')} \leqslant k'; \quad \frac{\alpha' \cdot \dfrac{1 - p(\alpha')}{1 - \alpha'}}{p(\alpha')} \leqslant k''; \quad \text{and} \quad \frac{\alpha' \cdot \dfrac{N + N_0}{2N}}{p(\alpha')} \leqslant k'''$$

in the last of which he assumes N_0 to be known.

The latter alternative is, however, rather unsatisfying because N_0 varies with α', which is to be determined. It is even possible to have $\tilde{N}_0 > N$.

When for example $k = 0.01$, these conventions mean that, on the average, at most 1 of 100 significances is false. If the number of tests is relatively small, for example $N = 10$, at most 0.1 false significances can be expected, but if $N = 1000$ this number increases to 10. (If, however, N increases, because we are including more treatments, it is probable that $N - N_0$, the number of tests where the alternative hypothesis is true, also increases.) These circumstances can give rise to a discussion concerning the criterion:

$$\frac{\alpha'}{p(\alpha')} \leqslant \frac{k}{N}$$

for keeping the absolute number of false significances down. Unfortunately for large N this would make the method too insensitive. For the sake of completeness we also relate the number of false significances to the total number of tests:

$$\frac{N \cdot \alpha'}{N} \leqslant \frac{k}{N}$$

which is Fisher's method discussed above.

In the following only the criterion $\alpha'/p(\alpha') \leqslant k$ will be studied. As $n(\alpha') = N \cdot p(\alpha')$, given α', is the only random variable, it can be worthwhile to re-phrase the criterion thus: If for some α'

$$n(\alpha') \geqslant \frac{N \cdot \alpha'}{k}$$

then there are $n(\alpha')$ "approved" significances. This procedure can be illustrated as in fig. III: 1.

Eklund's method is only used to determine the individual level of significance; it can be applied to N tests of any kind. One starts making the tests at the level $\alpha' = k$. If the criterion $n(\alpha') \geqslant N\alpha'/k$ is not satisfied, a lower value of α' is tried until the criterion is satisfied; that is, until the curve in fig. III: 1 touches the straight line. If this happens for $\alpha' = \alpha_0$, we say that we have obtained $n(\alpha_0)$ "mass-significances" ("approved" significances). It is theoretically possible that the curve touches the straight line more than once, but let us state that the method consists of choosing the largest α' for which this happens. Of course it is not possible to plot $n(\alpha')$ for every α'-value, but it is enough to study the values that make $N\alpha'/k$ an integer, namely:

$$\alpha' = k; \quad \frac{N-1}{N}k; \quad \frac{N-2}{N}k; \quad \ldots$$

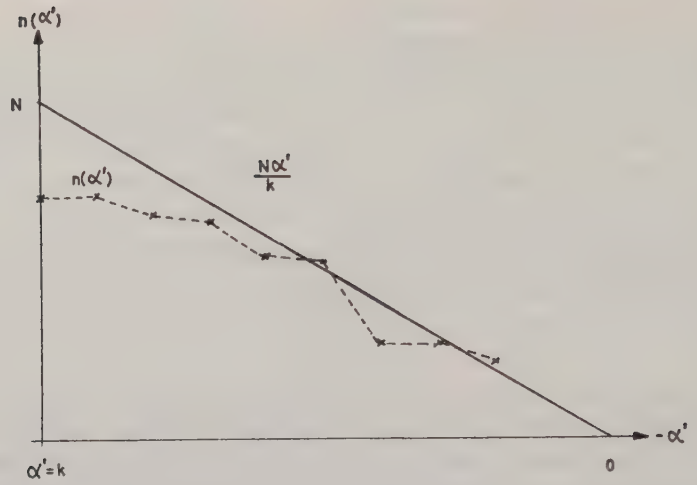


Fig. III: 1. The number of significances $n(\alpha')$ plotted against the individual level of significance α'.

In fact only few α'-values need be investigated, as $n(\alpha')$ can not increase with decreasing α'. After a certain $\alpha' = \alpha_1$, where $n(\alpha_1) < N\alpha_1/k$, it is easy to compute the largest α' which could give $n(\alpha_1) \geqslant n(\alpha') \geqslant N\alpha'/k$.

Eklund's method has its own definition of risks for type-I-errors, but a comparison with other methods would be facilitated by getting an idea of the probability α for at least one significant result when the null hypothesis is true for all the N tests.

The observed polygon $n(\alpha')$ can touch the straight line $n(\alpha') = N\alpha'/k$, for

$$\alpha' = k; \quad \frac{N-1}{N}k; \quad \frac{N-2}{N}k; \quad \ldots \frac{N-(N-1)}{N}k; \quad 0$$

Let us introduce $P_\nu; \nu = 0, 1, 2 \ldots N$; for the probability that $n(\alpha')$ touches the line for $\alpha' = \frac{N-\nu}{N} \cdot k$ but not for any larger value of α'. Such a "first touch" is bound to occur at one and only one of the points $\frac{N-\nu}{N} \cdot k$; $\nu = 0$, $1, 2 \ldots N$. Thus we have:

$$\sum_{\nu=0}^{N} P_\nu = 1$$

If there is a touch before $\nu = N$, false significances are obtained, (as we have assumed that the null hypothesis is true in every test); that is, the

total level of significance is:

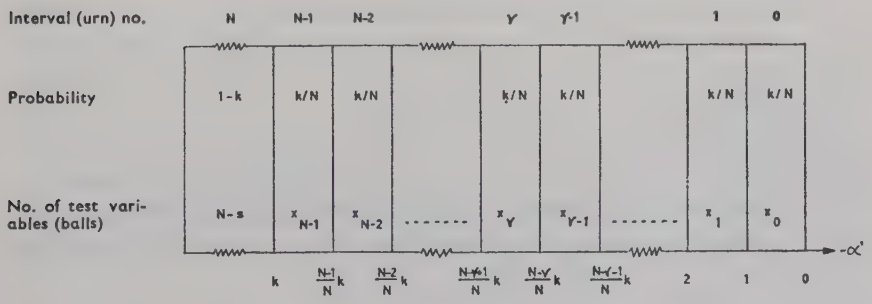
$$\alpha = \sum_{\nu=0}^{N-1} P_\nu = 1 - P_N$$

The line $n(\alpha') = \frac{N \cdot \alpha'}{k}$ is touched, for $\alpha' = \frac{N-\nu}{N} \cdot k$, if $N-\nu$ tests are significant at level $\frac{N-\nu}{N} \cdot k$. This touch is a first touch if no test variable has a value between the probability levels $\frac{N-\nu+1}{N} \cdot k$ and $\frac{N-\nu}{N} \cdot k$; at most one variable has a value in the interval $\frac{N-\nu+2}{N} \cdot k - \frac{N-\nu+1}{N} \cdot k$; at most two have values in the interval $\frac{N-\nu+3}{N} \cdot k - \frac{N-\nu+1}{N} \cdot k$; ...; at most $\nu-1$ have values in the interval $k - \frac{N-\nu+1}{N} \cdot k$. When the N tests are mutually independent the problem of computing P_N is equivalent to the following urn problem: N balls are thrown at random into $N+1$ urns, urn no. 0, 1, ..., $N-1$, N. At every throw each of the urns no. 0, 1, ... $N-1$ has the probability k/N, and urn no. N the probability $1-k$ of receiving the ball. Find the probability for the following distributions of balls among the urns:

$$x_0 = 0,\ x_1 \leqslant 1,\ x_1 + x_2 \leqslant 2,\ x_1 + x_2 + x_3 \leqslant 3,\ \ldots,\ x_1 + x_2 + \ldots + x_{N-1} \leqslant N-1$$

(See table III: 2.)

TABLE III: 2. *$N+1$ urns separated by the significance points $\frac{N-\nu}{N} \cdot k$; $\nu = 0, 1, \ldots N$. The probability of each individual urn of receiving a thrown ball is given, together with the number of balls in each urn.*



Numerical computations for small N, suggested that $P_N = 1 - k$ which can be proved to be generally true as follows:[1]

$$P_N = \left(\sum_{\nu=0}^{m} x_\nu \leqslant m \, ; \, m = 0, 1, 2 \ldots N \right)$$

$$= \underset{x_N}{E} \left\{ P\left(\sum_{\nu=0}^{m} x_\nu \leqslant m \, ; \, m = 0, 1, 2 \ldots N-1 \,\middle|\, x_N = N - s \right) \right\}$$

$$= \underset{s}{E} \left\{ P\left(\sum_{\nu=0}^{m} x_\nu \leqslant m \, ; \, m = 0, 1, 2 \ldots N-1 \,\middle|\, \sum_{\nu=0}^{N-1} x_\nu = s \right) \right\}$$

Let us introduce

$$\psi(s, N) = P\left(\sum_{\nu=0}^{m} x_\nu \leqslant m \, ; \, m = 0, 1, 2 \ldots N-1 \,\middle|\, \sum_{\nu=0}^{N-1} x_\nu = s \right)$$

When s balls are thrown independently into N equally probable receiving urns

$$\psi(s, N) = 1 - \frac{s}{N}; \; s = 0, 1, 2 \ldots N$$

which can be proved by induction in N. It is apparently true for $N = 2$ urns. If it is true for $N - 1$ urns, it is true for N as:

$$\psi(s, N) = \underset{x_{N-1}}{E} \{\psi(s - x_{N-1}; N-1)\} = \underset{x_{N-1}}{E} \left\{ 1 - \frac{s - x_{N-1}}{N-1} \right\} = 1 - \frac{s - (s/N)}{N-1} = 1 - \frac{s}{N}$$

Hence $$P\left(\sum_{\nu=0}^{m} x_\nu \leqslant m \, ; \, m = 0, 1, 2 \ldots N \right) = \underset{s}{E} \{\psi(s, N)\}$$

$$= \underset{s}{E} \left\{ 1 - \frac{s}{N} \right\} = 1 - \frac{N \cdot k}{N} = 1 - k$$

When there are positive correlations between some of the test variables k is an upper limit for the total level of significance α. In that case it would not be necessary to have Eklund's method preceded by an F-test, as the same protection against false significances as that test can give is already incorporated. However, Eklund's method cannot be said to have k as an upper limit for the error rate per experiment, because when some null hypotheses are not true, the number of false significances are allowed to increase. Thus the total level of significance will be larger than k.

With negatively correlated test variables α may also be larger than k.

[1] This proof is given by Docent Bengt Rosén, Uppsala.

Numerical illustrations.—First comes a more detailed study of the results of applying Eklund's method to one of the investigations of washing-machines referred to in (38). The number of machines was 9, and, as the interest lay in making pairwise comparisons, $N = \binom{9}{2} = 36$. With the help of comprehensive tables of the normal distribution (35) and an approximate connection between t- and λ-values:[1]

$$t_{\alpha;\nu} \approx \lambda_\alpha + (4\nu)^{-1}(\lambda_\alpha^3 + \lambda_\alpha)$$

the following is obtained:

α'	$n(\alpha')$
0.05	25
0.04	23
0.03	23
0.02	20
0.01	16
0.005	14
0.002	11

These results are illustrated graphically in fig. III: 2, where the straight lines $n(\alpha') = N\alpha'/k$ for $k = 0.01$ and 0.05 have also been drawn.

From fig. III: 2 one can see that with $k = 0.05$, Eklund's method gives 23 significances (the same as for Duncan's multiple range test with its $\alpha = 0.05$), with $k = 0.01$ it gives 11 significances (one more than Duncan's). Fisher's method with $\alpha \leqslant 0.072$ gives 11 significances.

Another two examples, in which the over-all significance was obvious, has been studied. In the first one, in which 9 washing-machines were compared in 60 blocks, the sign test was applied to all possible pairs. The following results were obtained:

	Fisher's method	Eklund's method
$\alpha' =$	$\alpha \leqslant$	$\frac{N\alpha'}{n(\alpha')}$
0.013488	0.486	0.0221
0.006216	0.224	0.0102
0.002669	0.096	0.0046

Thus Eklund's method with $k = 0.01$ and Fisher's with $\alpha \leqslant 0.096$ gave the same number of significances.

The second example, a psycho-physical experiment with 4 "treatments" and $N = 6$ pairwise tests, gave the following table:

[1] This equation is taken from (37), page 80.

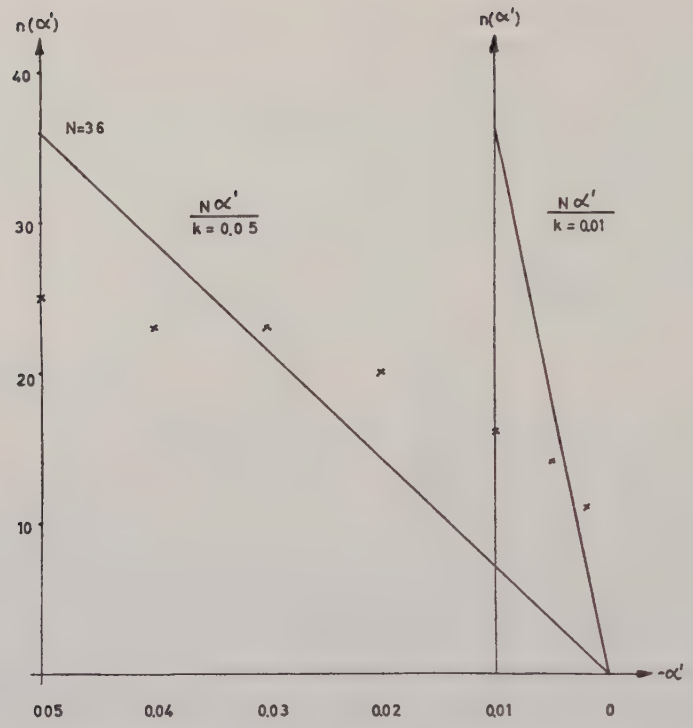


Fig. III: 2. The number of significances $n(\alpha')$ among the N tests plotted against the individual level of significance α'. The "significance line" $n(\alpha') = N\alpha'/k$ is drawn for $k = 0.05$ and $k = 0.01$.

$\alpha' =$	Fisher's method $\alpha \leqslant$	Eklund's method $\frac{N\alpha'}{n(\alpha')}$
0.02	0.12	0.06
0.01	0.06	0.03
0.002	0.012	0.012
0.001	0.006	$(n(\alpha') = 0)$

The result was 2 significances for Eklund's method with $k = 0.05$, for Fisher's with $\alpha \leqslant 0.06$, and for Duncan's with $\alpha = 0.05$.

Finally 3 examples with non-significant F-tests have been picked from different text books, and pairwise comparisons have been made with Eklund's, Duncan's and the multiple-t-method. The number of significances are given below:

Example no.	Number of treatments	Number of observations/ treatment	Number of significances with		
			Eklund's method	Duncan's method	Multiple t-method
1	5	5	0	0	2
2	7	3	0	0	1
3	10	10	0	3	3

Example no. 3, taken from (40), consisted of 10 random samples from one and the same normal population.

For small N-values there is an abundance of methods for making multiple comparisons, methods that give about the same results. In very broad investigations, if, for example, all pairwise comparisons between 56 treatments are desired, most of the methods become too insensitive. Eklund's method could then be easily applied as an alternative to the disputed new multiple range test of Duncan's.

CHAPTER 5

PRACTICAL RECOMMENDATIONS WITH FURTHER NUMERICAL ILLUSTRATIONS

The following recommendations may sometimes seem rather subjective, and it is true that they are not justified by a great number of mathematically proved facts. Instead they are based on what the author has been able to learn from discussions in the literature on the subject, from numerical examples carried out to find which tests are decent approximations, and from his own practical experience. These recommendations relate what the author would do faced with the problem in question.

As will be illustrated in this chapter, most of the methods for making multiple comparisons are very insensitive when the number of treatments in the experiment is large. Therefore separate recommendations are given for designs of a moderate size (5.1) and for very large designs (5.2).

5.1. Designs of a moderate size

5.1.1. When studying contrasts which involve more than two treatments, the contrast may, for example, consist of a difference between the mean of one group of treatment means and the mean of another group. Then the choice of a method of making multiple comparisons will principally lie between

the S-method,

Fisher's method, and

Eklund's method.

Dunn (13) has compared the lengths of the confidence intervals for the two first mentioned methods. Her result was, that if the number m of contrasts planned to be studied is not too large in comparison with a, the number of treatments, then Fisher's method gives shorter intervals. For example when $a=10$, and the number of degrees of freedom for the estimate of the variance, $\nu,=20$ m may be up to 300 without being "too large". (An experiment including 10 treatments could well be regarded as being of moderate size if, for example, we are interested in all $\binom{10}{2}$ paired

contrasts, but if it should be in our interest to study as many as 300 contrasts, the methods recommended for moderately large designs become insensitive.) Thus it seems as if Fisher's method would be the best of these two in practice, except possibly for contrasts that are not planned but suggested by the data. Scheffé (37) recommends Fisher's method for planned comparisons, but also that a part of the total level of significance should be spared for testing by the S-method on data-suggested comparisons.

Eklund's convention is more sensitive than any of the other two. For many of the numerical examples studied, Eklund's method with $k=0.01$ gave the same result as Fisher's with $\alpha=0.05$. But Eklund's method has its drawbacks which have been discussed earlier and besides, in Fisher's method α is an upper limit (experimentwise error rate) for the risk of committing an error of type I; hence α can be chosen somewhat larger than usual. Therefore it seems like the latter method is to be preferred. Consequently it is recommended to use Fisher's heightening of the significance level, and in addition possibly, to spare a fraction of α for testing by the S-method, which is to be used on non-planned contrasts. However, most or nearly all contrasts of interest can be planned without increasing m to such an extent that the S-method would give shorter intervals. Significant contrasts constructed after looking at the data do not often tell us anything, except possibly when they are constructed by examining broad exploratory investigations.

5.1.2. Sometimes it is of greatest interest to make all possible pairs of comparisons, as for example when we want to rank a number of treatments according to a certain characteristic. Several methods, for example Newman–Keuls', Duncan's and the T-method, are primarily intended for such comparisons and are for that type of contrasts more sensitive than other more general methods. The T-method for example has, for pairwise tests, a better power than the S-method, as has been pointed out in Scheffé's text book. But even Fisher's and Eklund's methods could of course be used for these situations. However, Eklund's and Duncan's methods will be avoided as long as there are other methods that can be used. There are several reasons for this: in Eklund's method the data determines the individual levels α', while Duncan's method is laborious and has questionable significance levels; like Newman–Keuls' method it does not give confidence intervals. The T-method and Fisher's method both give confidence intervals, and they have easily understandable definitions of significance level. The T-method is somewhat more sensitive but requires special tables and also demands that the means have equal variances. Again we recommend

Fisher's method, which permits the use of individual variances, that is variance-estimates from the contrast in question only. If there are strong reasons to believe that the assumptions of the T-method are satisfied even after the over-all null hypothesis is rejected, this method can of course be used to advantage. However the difference between the two methods need not be particularly large (see the example in paragraph 4.1).

5.1.3. In practice, experiments are often made which include a standard treatment, for example: a thoroughly investigated washing-machine, which is included in all the series of an experiment using sets of different machines, a standard tone compared with some of its variations, or a standard sort of wheat compared with some modified sorts. In such situations it would be most interesting to compare each treatment with the standard separately. It is desirable to know which experimental treatments are better than the standard, which are worse, and which are equivalent to the standard. For comparisons with a standard a special method has been worked out by Dunnett (15, 16). It would also be possible to use Fisher's method. For the same reason stated above, Eklund's method will not be discussed. Let us compare Dunnett's and Fisher's methods. The contrasts $\bar{x}_i - \bar{x}_0$; $i = 1, 2 \ldots p$, where $\bar{x}_0$ is the mean of the standard treatment, will apparently be positively correlated. As Dunnett has shown the correlation is

$$\varrho_{ij} = \frac{1}{\sqrt{(r_0/r_i + 1)(r_0/r_j + 1)}}$$

and he has computed his tables by assuming the correlation $\frac{1}{2}$, which is true when $r_i = r_0$ for all i's that is, when the number of observations are constant from treatment to treatment. In his later paper corrections are given which are to be used when the standard includes more observations than the other treatments. Fisher's method (based on Bonferroni's inequality) reckons with the most unfavourable case—some negative correlation, and is therefore somewhat more insensitive. Dunnett's old tables, which are now reprinted in text books and probably used a good deal, give only upper limits for the significance level in double-sided tests, and they are therefore of only slightly better use than Fisher's method. In his new table, however, Dunnett has removed this drawback. The critical values for the two methods have been compared for $\alpha = 0.05$, and $4(p+1)$ degrees of freedom in table III: 3.

For moderately extensive experiments the difference is not important even if Dunnett's new tables are used, and if the level is increased somewhat, as it often is using Fisher's method, the methods will be practically

TABLE III: 3. *Critical values.*

p	D.f.	For Fisher's method	From Dunnett's old table	From Dunnett's new table
1	8	2.31	2.31	2.31
2	12	2.52 ($\alpha = 0.0500$) 2.50 ($\alpha = 0.0520$)	2.50	2.50
3	16	2.64 ($\alpha = 0.0504$) 2.59 ($\alpha = 0.0558$)	2.63	2.59
4	20	2.72 ($\alpha = 0.0504$) 2.65 ($\alpha = 0.0584$)	2.70	2.65
5	24	2.78 ($\alpha = 0.0500$) 2.70 ($\alpha = 0.0600$)	2.76	2.70
9	40	2.92 ($\alpha = 0.0504$) 2.80 ($\alpha = 0.0684$)	2.90	2.81

equivalent. Fisher's method is to be the recommended because of the higher accessibility to its tables[1] and the greater flexibility of its use; for example, it is possible to use any level. The merit of using very few methods should not be underestimated either.

5.1.4. The standard assumptions made in the analysis of variance and, in consequence, even in the following multiple comparisons are: normal distributions, equal variances, and uncorrelated residuals. The last mentioned assumption will not be treated here, because correlated residuals may have serious consequences for the analysis and because it is, as a rule, possible to plan the experiment so that this assumption is satisfied.[2] However, in practice one must often reckon with deviations from normality and variance equality. Suitable methods in some such cases will be discussed below.

a) *Completely randomized one-way designs*

Quantitative variation.—If the variances are equal and the central limit theorem is applied to the contrast, it is well known that rather few observations per treatment are needed to give a good enough normal approxima-

[1] Dunnett's new tables give critical values for $\alpha = 0.05$ and 0.01, $p = 1$ (1) 12, 15 and 20, $\nu = 5$ (1) 20, 24, 30, 40, 60, 120, ∞. Fisher's method with t-tests requires, for small m's, ordinary t-tables; otherwise Pearson-Hartley's tables (35) can be used for almost any values of α, p and ν.

[2] At least in those experiments, that are considered here, and are not of the time-series type. Some cases with correlated residuals will be considered in part IV.

tion. See, for example, van der Waerden (49), who seems to imply that 4 observations per treatment are enough for most of the distributions met in practice, at least for the over-all F-test. But even t-tests of difference between only two means could probably be used in rather small experiments. When the variances are equal, it would also be possible to use rank tests. It is true that they assume identical distributions, but equal means (the null hypothesis) and equal variances ought to give a good enough approximation. However, as the rank tests are insensitive when applied to small groups, at least compared to the t-test, it is probably very difficult to reach significant results when using them in experiments so limited that the t-tests can not be used. Thus it is improbable that deviations from normality when the variances are equal, have any importance unless the number of observations per treatment is very small. When this is not the case t-tests with Fisher's heightening of the significance level can be used in making multiple comparisons.

However, multiple comparisons are preceded by an F-test which rejects the hypothesis that all means are equal. When the means vary it is often wise to reckon with varying variances too. Then, for every contrast only the variances for the treatments included in the contrast will be used (individual variances). For large samples, the occurrence of differences between these variances is of fairly little importance, especially when contrasts in pairs of treatments are considered. (See, for example, Scheffé's text book.) When the populations are normal the t-tests will be rather good approximations even for small samples. Scheffé (37) presents some numerical results, among them, results from two samples with 7 observations each. If the variances are not extremely divergent, the approximation to the risk of committing an error of the first kind will hold for still fewer observations. This presupposes one important assumption: that the treatment-groups are equally large; a very uneven distribution of the observations may seriously disturb the test. In experiments where there is a risk of missing observations, we have to make sure at the start that we shall have large samples left over for the analysis.

When the populations are not normal, the variances very divergent, and the samples small, it is sometimes possible to use data transformations in order to make their distributions more like normal distributions and in order to give these populations less divergent variances. In this case something must be known about the original distributions. Another possibility might be to use rank tests, but they can not be recommended when the variances differ greatly.

Qualitative variation.—When the data only consists of rank orders there are special rank tests by Steel (44) and Dunn (14), the last one using Fisher's heightened significance levels, which have been presented earlier. As Dunn points out it seems most reasonable to use Steel's method in limited experiments with equally large groups, as we then have access to Steel's tables, but to use Dunn's procedure I when large samples are investigated. For a discussion of the hypothesis for these tests and of the possibility of using F- and t-tests on ranks, see part IV chapter 3.

b) *Randomized block designs with one factor*

Quantitative variation.—When the distributions are the same in every block, one can use individual t-tests with heightened levels, as long as the design does not include too few blocks. Generally speaking we cannot say which is the smallest number of blocks which allows us to apply the t-test. However, in many practical cases it is possible to reduce their number considerably, perhaps down as far as between 8 to 10 blocks, as there remains at least the difference between two treatments for every block. As usual we consider a random model, that is the blocks are considered sampled. If the model is fixed an interaction between treatments and blocks would decrease the power of the test. (Compare what has been said about tests with mixed model in part II, paragraph 3.2. The problem of asymmetry met with there will disappear when we consider contrasts in pairs of treatments.) It has been previously mentioned that sign tests are preferred to individual t-tests, only if the distributions vary from block to block. In practice this will probably be the case only when the blocks are experiments executed at different times and perhaps in different places.

Qualitative variation.—When the data consists of a set of ranks for every block the sign test with Fisher's heightened level for pairwise comparisons can be used. Even more complicated contrasts could, in principle, be studied with the sign test, but when ranks are considered there is a risk of obtaining many ties. It will be shown in part IV chapter 3, that an increase in the probability of the occurrence of ties will decrease the power rapidly. For example, if we are studying the contrast $2 \cdot A_1 - A_2 - A_3$, and if we rank only within the contrast, then the probability of a tie is as high as $\frac{1}{3}$ when the 6 rank orders are equally probable. However, if we rank all treatments in the block and use these rank numbers for any contrast, the "tie-probability" may decrease. If all rank orders between the four treatments $A_1 \ldots A_4$ are equally probable, the probability of obtaining a tie

for the contrast $2 \cdot A_1 - A_2 - A_3$ is only $\frac{1}{6}$. In large designs with many blocks it is also possible to use t-tests on ranks, but the hypothesis is still qualitative as it is for the sign test. (Compare part IV, chapter 3.)

c) *Designs with more than one factor (Factorials)*

Our study concerns only one factor of interest (if we are interested in all combinations of several factors it is equivalent to a one-factor problem); other factors are regarded as classification factors only—as a new block type. Thus from this aspect factorials imply nothing new when compared to randomized blocks with one factor. If we have a block factor with b levels and a classification factor with c levels, we can say equally well that we have $b \cdot c$ blocks. However, if the classification factor is nested within the factor of special interest, we have to take the mean of the c levels and reckon with only b blocks. (Compare the meteorological example in part IV, chapter 3.)

5.2. Very large designs

When the investigation includes a very large number of treatments, for example, when watches of 42 different makes are tested or 56 tones are compared in a psycho-physical experiment, all methods described here for making multiple comparisons, except Duncan's and Eklund's, are too insensitive to be used. Though Eklund's and Duncan's methods are not particularly satisfying, they seem to be the only usable ones, unless the LSD-method be accepted. Some practical difficulties arising in Duncan's method could be removed by combining it with Fisher's method, in such a way that the individual levels: $\alpha' = \alpha/(a-1)$ come into use; where a is the total number of treatments, even when the number of tests to be performed is larger than $a-1$. However, we can only make this simplification at the cost of a loss in power.

These circumstances are illustrated below by both of the above mentioned examples.

Example 1. Testing watches

In an investigation of watches, one of the properties of the watches studied was their regularity of functioning. By this was meant their tendency to gain (lose) the same number of seconds every day, and it was measured by the variance s^2 of the gains and losses on consecutive days. Forty-two different makes, each represented by 3 watches were compared. Two makes were rejected because they were represented by 4 and 2 watches respectively in the primary data. With only 3 observations per group it is

TABLE III: 4. *Means of logarithmic variances for 3 watches each of 38 tested makes.*

Make no.	$\bar{x}=$	Make no.	$\bar{x}=$
35	3.3644	38	2.3421
36	3.2583	29	2.3263
2	3.2157	22	2.2795
20	3.1869	14	2.2733
21	3.0758	12	2.2357
34	3.0598	27	2.2010
37	3.0532	23	2.0674
24	2.9336	28	2.0652
26	2.9104	4	1.8509
18	2.7356	33	1.7798
39	2.6620	5	1.7471
32	2.5568	10	1.7128
13	2.4971	1	1.6526
11	2.4887	16	1.5899
40	2.4678	7	1.5326
6	2.4455	8	1.4439
9	2.4410	15	1.4222
3	2.3710	25	1.0431
31	2.3510	17	0.9835

particularly important that the variances be equal, so that a common estimate with many degrees of freedom may be used. In the hope of obtaining better homogeneity of variances and perhaps a better approximation of the normal distribution logarithms were taken. Thereafter two more makes were rejected, one because its resulting variability was extremely large compared to that of the others', and the other because of its extremely small variability. Thus 38 makes were left for comparison. The 38 means are given in the order of their magnitude in table III: 4.[1]

The following estimates of the standard deviations between watches of the same make were obtained:

$$\hat{\sigma} = 0.4195$$

$$\hat{\sigma}_{\bar{x}} = 0.2423$$

$$\hat{\sigma}_{\bar{x}_i - \bar{x}_j} = 0.3425$$

The S-method for pairwise comparisons at the level 5% gives

$$S \cdot \hat{\sigma}_{\bar{x}_i - \bar{x}_j} = \sqrt{(38-1) \cdot 1.56} \cdot 0.3425 = 2.6026$$

[1] The relevance of the obtained measurements and of the transformation can be discussed. In this paper we disregard this, consider the observations as given, and use them only as illustrations of the statistical methods.

As this is larger than the difference between makes no. 35 and 17 no significant result is obtained in spite of a very strongly significant F-test.

Eklund's method. As we want to rank the makes, we must test all pairs. This means that $N = \binom{38}{2} = 703$.

If $k = 0.05$, we obtain:

α'	$t_{\alpha'}$	$t_{\alpha'} \cdot \hat{\sigma}_{\bar{x}_i - \bar{x}_j}$	$n(\alpha')$	$\frac{N\alpha'}{n(\alpha')}$	Next $\alpha' \leqslant$
0.05	1.99	0.6816	339	0.1037	0.0241
0.03	2.2109	0.7572	288	0.0732	0.0205
0.02	2.377	0.8141	258	0.0545	0.0183
0.018	2.4169	0.8278	252	0.0502	—

that is 252 "mass-significances"

$k = 0.01$ gives:

α'	$t_{\alpha'}$	$t_{\alpha'} \cdot \hat{\sigma}_{\bar{x}_i - \bar{x}_j}$	$n(\alpha')$	$\frac{N\alpha'}{n(\alpha')}$	Next $\alpha' \leqslant$
0.01	2.643	0.9052	214	0.0329	0.00304
0.003	3.0716	1.0520	160	0.0132	0.00228
0.002	3.200	1.0960	148	0.0095	—

Perhaps it would have been possible, for $k = 0.01$, to reach a greater number of "mass-significances", when $\alpha = 0.0022$ or 0.0021, than the obtained 148, but it is difficult to find tables sufficiently detailed. In any case the method possesses a great enough power.

Duncan's method. This method requires only one level of significance, but, on the other hand, as many as 38-1 different critical values. Existing tables are not this comprehensive; one has to interpolate in them. The critical values are given in table III: 5.

For comparison we repeat Eklund's critical values

$k = 0.05$	$k = 0.01$
0.8278	1.0960

Duncan's method gives 251 and 150 significant pairs for $\alpha = 0.05$ and 0.01 respectively, Eklund's 252 and 148 for the same k-values respectively. The difference between the results of the two methods for the level 5%, is, for make no. 25, that Eklund's method gives one significance less than Duncan's, and, for makes no. 1 and 12, that Eklund's method gives one

TABLE III: 5. *Critical values, s_d, for Duncan's method, with p treatments and a level of significance α.*

$\alpha = 0.05$				$\alpha = 0.01$			
p	s_d	p	s_d	p	s_d	p	s_d
2	0.6835	21	0.8419	2	0.9103	21	1.0965
3	0.7191	22	0.8437	3	0.9462	22	1.0993
4	0.7431	23	0.8451	4	0.9723	23	1.1019
5	0.7599	24	0.8466	5	0.9917	24	1.1044
6	0.7732	25	0.8478	6	1.0070	25	1.1066
7	0.7838	26	0.8490	7	1.0196	26	1.1088
8	0.7926	27	0.8500	8	1.0303	27	1.1107
9	0.8001	28	0.8510	9	1.0392	28	1.1126
10	0.8064	29	0.8519	10	1.0472	29	1.1143
11	0.8117	30	0.8529	11	1.0540	30	1.1160
12	0.8166	31	0.8536	12	1.0603	31	1.1177
13	0.8207	32	0.8543	13	1.0659	32	1.1194
14	0.8245	33	0.8550	14	1.0710	33	1.1209
15	0.8277	34	0.8556	15	1.0756	34	1.1223
16	0.8308	35	0.8562	16	1.0799	35	1.1237
17	0.8335	36	0.8568	17	1.0838	36	1.1250
18	0.8359	37	0.8571	18	1.0872	37	1.1262
19	0.8381	38	0.8575	19	1.0906	38	1.1274
20	0.8401			20	1.0937		

more significance than Duncan's. However, for make no. 12 the difference in question is exactly equal, up to four decimal places, to the critical value for Duncan's method, and with less errors of rounding off might have given significance too. At the 1% level, Eklund's method gives two significances less for the best make, no. 17.

The agreement between the two methods must be said to be very good in this example.

Example 2. A psycho-physical experiment

The perception of pause length between tones of different durations, intensities and frequencies was studied in this experiment (24). The subjects were asked to adjust the pause lengths with the aid of various wheels. They were to make the length of these pauses as near to the length of a certain standard pause as they could. This was done for 56 different tonal conditions. Seven subjects made 4 adjustments[1] each for every condition, that is in all 224 adjustments per subject. The order between these 224 adjustments was independently randomized for every subject.

[1] For some reason these were averaged two and two, and thus only two values per tone and subject were used in the analysis.

TABLE III: 6. *Means of 2 observations of the perception of pause length for different tonal conditions.*

Condition no.	$\bar{x}$	Condition no.	$\bar{x}$
1	62.1	29	85.4
2	66.1	30	85.6
3	67.1	31	85.6
4	67.7	32	85.6
5	70.1	33	85.9
6	70.1	34	86.6
7	70.1	35	86.8
8	70.3	36	87.5
9	73.1	37	88.3
10	74.9	38	89.0
11	75.6	39	92.0
12	75.6	40	92.7
13	75.7	41	93.8
14	76.6	42	93.9
15	76.8	43	94.1
16	77.5	44	97.0
17	77.9	45	97.2
18	78.6	46	98.1
19	79.1	47	98.6
20	79.5	48	99.6
21	80.1	49	100.7
22	81.2	50	101.1
23	81.2	51	102.1
24	81.3	52	105.9
25	81.7	53	106.6
26	81.9	54	106.9
27	84.2	55	117.1
28	84.3	56	118.4

The observed means per condition (in milliseconds), ranked from the smallest to the largest value, are given in table III: 6.

An analysis of variance of the primary data gave the following table:

Sources of variation	D.f.	MS
Conditions	55	2 258.06
Subjects	6	19 970.43
Interaction	330	422.51
Residual	392	346.54

Estimates of error: $\hat{\sigma}_{\bar{x}} = 5.4936$; $\hat{\sigma}_{\bar{x}_i - \bar{x}_j} = \hat{\sigma}_d = 7.769$.

The hypothesis that the pauses are perceived as equally long for all 56 conditions (averaged over the subjects) is tested with the quotient:

$$F = \frac{2258.06}{422.51} = 5.34$$

The mean square for interaction is used in the denominator because we want to generalize the results to a greater group of individuals, viz. the group of musicians of which the 7 subjects can be considered as representative. The critical value when $\alpha = 0.0005$, is close to 2.0; the hypothesis is therefore rejected.

To find out which mean values differ, we can try the S-method with $\alpha = 0.05$. The following is then obtained:

$$\hat{D}_S = 7.769 \cdot \sqrt{55 \cdot 1.4} = 68.17$$

The difference between the largest and smallest of the observed means is: $118.4 - 62.1 = 58.3$; that is no pair of treatments can be declared significant using the S-method at the 5% level, in spite of the very strong significance obtained by the F-test. Fisher's method, applied to all $\binom{56}{2} = 1540$ pairs, would use the individual level $\alpha' = \alpha/1540$. When for example $\alpha = 0.077$, this means that $\alpha' = 0.00005$. Apparently none of these methods can be used for making all pairwise comparisons between a very large number of treatments.

As the 1540 tests for pairs are equivalent to 55 independent tests, one could use, as Duncan might have done, the individual level $\alpha/55$, for example, $0.055/55 = 0.001$. Using this variant the same result is obtained as by Duncan's method at the 1% level. The results from Eklund's and Duncan's methods both at the 5% level are given below:

Eklund's method with $k = 0.05$.

α'	$t_{\alpha'}$	$t_{\alpha'} \cdot \hat{\sigma}_d$	$n(\alpha')$	$\frac{N\alpha'}{n(\alpha')}$	Next $\alpha' \leqslant$
0.05	1.9672	15.28	624	0.1234	0.0203
0.02	2.3376	18.16	487	0.0632	0.0158
0.014	2.4704	19.19	442	0.0488	—

Duncan's method with $\alpha = 0.05$.

This method was difficult to use with $\nu = 330$ degrees of freedom, as Harter's tables (27) include $\nu = 120$ and ∞, but nothing in between. In this paper we have chosen the ν which gives the largest critical value, s_d, that is, $\nu = 120$ when $p \leqslant 20$, and $\nu = \infty$ when $p \geqslant 20$. There is therefore a risk of underestimating the number of significances that would have been obtained when $\nu = 330$. Moreover, a great number of critical values have been interpolated.

TABLE III: 7. *Critical values for Duncan's multiple range test.*

p	s_d	p	s_d	p	s_d
2	15.38	21	19.10[a]	40	19.79
3	16.19	22	19.15	41	19.81[a]
4	16.73	23	19.20	42	19.83[a]
5	17.12	24	19.26	43	19.85[a]
6	17.43	25	19.30[a]	44	19.87[a]
7	17.67	26	19.35	45	19.90[a]
8	17.88	27	19.39[a]	46	19.92[a]
9	18.06	28	19.43	47	19.94[a]
10	18.21	29	19.44[a]	48	19.96[a]
11	18.33	30	19.50	49	19.98[a]
12	18.45	31	19.54	50	20.00
13	18.55	32	19.57	51	20.01[a]
14	18.65	33	19.60[a]	52	20.03[a]
15	18.73	34	19.63	53	20.04[a]
16	18.80	35	19.66[a]	54	20.06[a]
17	18.87	36	19.69	55	20.07[a]
18	18.93	37	19.72[a]	56	20.09[a]
19	18.99	38	19.74		
20	19.04	39	19.77[a]		

[a] Linearly interpolated.

The total of 438 significantly different pairs were obtained, and, as this may be an underestimate, it must be said to be in very good agreement with Eklund's total of 442. However, the difference is, in fact, somewhat greater. Eklund's constant critical value 19.19 is larger than those of Duncan's used in testing the range of a limited number of means, but smaller than those used in testing the range of a large number of means (see table III: 7). Thus it is easier to detect significant differences between treatments that differ only a little in rank, using Duncan's method; while Eklund's method facilitates detection of significant differences between treatments, whose difference in rank is large. In table III: 8 the differences are shown in greater detail.

The above example has been used in this way in order to illustrate some methods in making multiple comparisons. In fact condition no. 18 was used as a standard, and the main interest was to compare it with each of the other conditions. If these 55 comparisons are made by using the above analysis, Eklund's and Duncan's methods render the following results:

no. 18 < 11 conditions

no. 18 > 0 conditions

In the above analysis N was equal to 1540 using Eklund's method, which, however, is a more flexible method and can be applied directly to the 55

TABLE III: 8. *Differences in numbers of significances found by Eklund's and Duncan's methods. Condition no. 1 is found to be significantly "less" than 33 conditions using Eklund's method and significantly less than 32 conditions by Duncan's method etc.*

Condition no.	Eklund's method		Duncan's method
1 <	33		32
2 <	28		24
3 <	23		22
4–9 <		The same number	
10 <	14		13
11–16 <		The same number	
17 <	12		11
18–25 <		The same number	
26 <	7		6
27–34 <		The same number	
35 <	4		5
36 <	3		4
37–45 <		The same number	
46 <	1		2
47 <	1		2
48 <	0		1
Total	442		438

comparisons, that is when $N = 55$. The result is then:

$$\text{no. } 18 < 9 \text{ conditions}$$

$$\text{no. } 18 > 0 \text{ conditions}$$

which is an example of what has been said before, viz. that by including some more tests in Eklund's method, we may increase the number of significances in the original tests.

Fisher's method ($\alpha \leqslant 5.5\%$) applied to these 55 tests, gives the same result as Duncan's at the 1% level. It may be interesting to note that Fisher's method gives more significances using individual variances than when a common variance is used. Thus the use of a common variance with many more degrees of freedom does not necessarily increase the power, if the variances are different.

In both these examples Eklund's and Duncan's methods have given essentially the same results for pairwise comparisons. Eklund's method is somewhat simpler and more flexible; it allows, for example, the use of individual variances and even gives simultaneous confidence intervals. However, the individual significance levels of this method are determined not only by the number of tests, as, for example, in Fisher's method, but

also by the observations themselves. This latter characteristic means that including some very bad or very good treatments may increase the number of significances that exist between the original treatments.

The problem of multiple comparisons with very large designs has been rather neglected. No solution that is satisfactory both theoretically and practically exists. It would be best to have one method that could be used for limited as well as extensive designs equally well. To-day Duncan's method comes nearest to this ideal, but it is not at all generally accepted; it is also very difficult to use in practice. If it is to be used, it ought to be preceded by an F-test. This will reduce the risk of finding significances when all a means are equal, but if one population mean differs very much from the others, the others being equal, there still exists the great risk of finding significances between the $a-1$ equal ones.

5.3. Summary

In this chapter Fisher's method has been recommended most of all, although the individual tests may be t-tests, sign tests or rank sum tests, and may have common or individual variance estimates depending on the assumptions. To follow this recommendation means that we plan a number, say N, of interesting comparisons and carry these through at a level of significance of α/N. Possibly a fraction α_0 could be spared for making comparisons not planned but suggested by the data, while the level $(\alpha-\alpha_0)/N$ would then be used for the planned comparisons. If it is clear that the classical assumptions are really satisfied, and we want to make all pairwise comparisons, then the T-method will give a somewhat higher power than Fisher's method.

As it has been recommended above to use methods in making multiple comparisons only after a significant over-all F-test, not only inequality between means but also between variances should be taken into consideration. As a rule it is therefore better to estimate the variance by only considering the treatments included in the contrast. This will require much numerical work. It would therefore be worthwhile to set up a computer program that would calculate all possible t-values using individual or common variances for differences between two treatments. The planned comparisons are then put forth, the confidence (significance) levels determined in proportion to the number of comparisons chosen, and the corresponding confidence limits easily found. It is also possible to synthesize t-values to give more complicated contrasts than pairwise contrasts, even if it would require some additional computations.

In very extensive designs Fisher's method becomes too insensitive. It has been suggested (following Duncan's example) that the significance level $\alpha/(a-1)$ be used, where a is the number of treatments in the experiment, as long as $a-1<N$. Duncan's and Eklund's methods have also been described, but no definite recommendations for or against have been given. To get rid of some of the practical difficulties arising in Eklund's method, a computer program for Fisher's method could easily be extended.

REFERENCES

1. Adams, J. K.: Basic statistical concepts. McGraw Hill, New York (1955).
2. Birnbaum, A.: On the analysis of factorial experiments without replication. Contributed paper at Annual Meeting of the institute of Mathematical Statistics, Cambridge Massachusetts, August 25–28 (1958).
3. Brownlee, K. A.: Statistical theory and methodology in science and engineering. Wiley, New York (1960).
4. Daniel, C.: Fractional replication in industrial research. Third Berkeley Symposium on Mathematical Statistics and Probability, Vol. V, 87–98. University of California Press (1956).
5. Davies, O. (ed.): The design and analysis of industrial experiments. Hafner Publishing Co., New York (1956).
6. Dixon, W. J. and A. M. Mood: The statistical sign test. Jasa 41 (1946), 557–566.
7. Duncan, D. B.: Significance tests for differences between ranked variates drawn from normal populations. Iowa State College thesis, (1947).
8. — A significance test for differences between ranked treatments in an analysis of variance. The Virginia journal of science 2 (1951), 171–189.
9. — On the properties of the multiple comparisons test. The Virginia journal of science 3 (1952), 49–67.
10. — Multiple range and multiple F-tests. Biometrics 11 (1955), 1–42.
11. — Multiple range tests for correlated and heteroscedastic means. Biometrics 13 (1957), 164–176.
12. Dunn, O. J.: Confidence intervals for the means of dependent, normally distributed variables. Jasa 54 (1959), 613–621.
13. — Multiple comparisons among means. Jasa 56 (1961), 52–64.
14. — Multiple comparisons using rank sums. Technometrics 6 (1964), 241–252.
15. Dunnett, C. W.: A multiple comparison procedure for comparing several treatments with a control. Jasa 50 (1955), 1096–1121.
16. — New tables for multiple comparisons with a control. Biometrics 20 (1964).
17. Eklund, G.: Massignifikansproblemet. Unpublished seminar paper. Uppsala (1962).
18. Eklund, G. and P. Seeger: Massignifikansanalys. Statistisk Tidskrift 3:e följden årg. 3 (1965), 355–365.
19. Federer, W. T.: Experimental design. Macmillan Co. N.Y. (1955).
20. Finney, D. J.: The joint distribution of variance ratios based on a common error mean square. Annals of Eugenics 11 (1941), 136–140.
21. Fisher, R. A.: The correlation between relatives on the supposition of Mendelian inheritance. Trans. Roy. Soc. Edinburgh, 52 (1918), 399–433.
22. — The design of experiments. Oliver and Boyd, Edinburgh (1935).
23. Gabriel, K. R.: A procedure for testing the homogeneity of all sets of means in an analysis of variance. Biometrics 20 (1964).

24. Gabrielsson, A.: Betoningsupplevelse i korta tonsekvenser. Planering och förundersökningar. Unpublished lic. thesis, Uppsala: Inst. of psychol. (1964).
25. Guenther, W. C.: Analysis of variance. Prentice-Hall, Inc., Englewood, Cliffs, N.J. (1964).
26. Harter, H. L.: Error rates and sample sizes for range tests in multiple comparisons. Biometrics 13 (1957), 511–536.
27. — Critical values for Duncan's new multiple range test. Biometrics 16 (1960), 671–685.
28. — Corrected error rates for Duncan's new multiple range test. Biometrics 17 (1961), 321–324.
29. Kendall, M. G.: The advanced theory of statistics. Ch. Griffin o Co. London (1948).
30. Keuls, M.: The use of the studentized range in connection with an analysis of variance. Euphytica 1 (1952), 112–122.
31. Kimball, A. W.: On dependent tests of significance in the analysis of variance. AMS 22 (1951), 600–602.
32. Kramer, C. Y.: Extension of multiple range tests to group means with unequal numbers of replications. Biometrics 12 (1956), 307–310.
33. — Extension of multiple range tests to group correlated adjusted means. Biometrics 13 (1957), 13–18.
34. Newman, D.: The distribution of range in samples from a normal population, expressed in terms of an independent estimate of standard deviation. Biometrika 31 (1939), 20–30.
35. Pearson, E. S. and H. O. Hartley: Biometrika tables for statisticians, Vol. I. Cambridge University Press (1956).
36. Scheffé, H.: A method for judging all contrasts in the analysis of variance. Biometrika 40 (1953), 87–104.
37. — The analysis of variance. Wiley N.Y. (1959).
38. Seeger, P.: Olika modeller för variansanalys av ett autentiskt exempel. Unpublished lic. thesis, Uppsala (1962).
39. Siotani, M.: Interval estimation for linear combinations of means. Jasa 59 (1964), 1141–1164.
40. Snedecor, G.: Statistical methods, 5th edition, page 245. Iowa state College Press (1956).
41. Steel, R. G. D.: A multiple comparison rank sum test: Treatments versus control. Biometrics 15 (1959), 560–572.
42. — A multiple comparison sign test: Treatments versus control. Jasa 54 (1959), 767–775.
43. — Answer to query 163. Biometrics 17 (1961), 326–328.
44. — A rank sum test for comparing all pairs of treatments. Technometrics 2 (1960), 197–208.
45. Tippett, L. H. C.: The method of statistics. Williams and Norgate, London (1931).
46. Tukey, J. W.: Comparing individual means in the analysis of variance. Biometrics 5 (1949), 99–114.

47. — Quick and dirty methods in statistics, part II, simple analyses for standard designs, Proceedings Fifth Annual Convention, American Society for Quality Control (1951), 189–197.
48. — The problem of multiple comparisons. Unpublished dittoed manuscript. Princeton University (1953).
49. VAN DER WAERDEN, B. L.: Mathematische Statistik. Springer-Verlag Berlin (1957).

Part IV

SCALE PROBLEMS

CHAPTER 1

INTRODUCTION

When we say that the assumptions made before using t- and F-tests are not satisfied, we usually mean that the distributions are not normal, that the variances are not equal or that the observations are not independent. Special deviations from the assumptions may be caused by using measurements at a weak level. The mean temperature at some place measured by controlled thermometers has a definite meaning to us, and it can be compared with the mean at some other place measured by other controlled thermometers. However, if someone judges the quality of a TV-picture according to a scale graded from 1–6 or the cleanness of glasses after they have been washed in a dish-washing machine according to a scale graded from 1–4, the measurements are definitely at a weaker level than when measuring temperatures. Ranking, for example, when someone is asked to rank towels in order of softness, constitutes another example of qualitative data of this kind.

Qualitative data are often made quantitative, and then parametric tests such as the t-test are applied to the resulting scores. Whether such operations are allowable or lead to gross errors depends on the type of data. Strictly speaking all observations are qualitative, but the possibility of transforming observations to numbers in a meaningful way varies.

The parametric tests are considered less suitable to be applied to qualitative data, but not only because the assumptions of normality and equal variances are seldom satisfied. It is more important that ordinary means and variances are considered meaningless when the measurements are at a weak level. Of course, the means and variances do not signify the same thing as when the measurements are quantitative, but, if one can find out what significance they do have and if it is relevant, the sole fact that the measurements are at a weak level should not prevent us from using the parametric tests.

When using scales of measurement of the type discussed above there exist a lot of non-parametric tests, some of which will be discussed in the following chapters. These tests are characterized by the fact that they examine other hypotheses than the parametric tests do in the quantitative

case, and also by the fact that the definition of additivity (no interaction) is different. When in some cases the t- or F-tests may be applied, for example, to ranks when the samples are very large, they also test the qualitative hypotheses. These hypotheses are often very vaguely formulated in text books.

In some border cases it may be difficult to decide which is the appropriate technique; for example, if the measuring scale varies; viz. where one unit of measurement at one end of the scale[1] is not equivalent with one unit at the other end. Such a scale obviously gives us better information than pure ranking can, but do means and standard deviations calculated from this information tell us the same thing as when the scale is constant? Apparently the hypothesis tested will be a bit changed, at least when there is interaction. Such a varying scale may arise for example when the scores are quotients between primary observations. In some experiments on washing-machines (17) the score was:

$$R = \frac{E - F}{J - F}$$

where E, F and J were accurate values of whiteness measured by a reflectometer before the washing (F), after a first washing (E) and after a common washing with garments from all machines (J). This score is very sensitive to observational errors when the denominator is near zero.

In this part tests of hypotheses for some qualitative situations will be discussed. Sometimes it is possible by trial-and-error methods to make interval estimates of parameters, but estimation will not be treated here. When data are qualitative there is often nothing to estimate. For example when towels are ranked according to softness and we cannot measure any magnitude of softness; what should be estimated? The following discussion will be concentrated on models, qualitative and quantitative hypotheses, but also on a number of developements of existing tests. Two scales will be treated: the nominal scale in chapter 2 and the ordinal scale in chapter 3.

[1] Note that scale is used in two different meanings here: Different scales of measurement, for example nominal scale, ordinal scale, and uniform or non-uniform scale on a coordinate axis.

CHAPTER 2

NOMINAL SCALE

2.1. Concerning the scale

This scale[1] is sometimes also called the classificatory scale. Objects, persons or characteristics are classified in different categories. For example, when studying hair colour observations can be classified as black, blond, red and so on. The data from an investigation on the nominal scale will be frequencies for the different categories. We will in this section treat the case of only two categories: one with objects having a certain characteristic, the other with objects not having it. Another restriction is that we consider only related samples, as it is the most interesting case. For a discussion of independent samples see (18).

2.2. Several related samples (randomized block designs)

2.2.1. *Introduction*

When the scale is nominal and one wants to test the hypothesis that a independent samples are drawn from populations with a common distribution one may use the ordinary χ^2-test. The following situation illustrates this. From each one of the three social classes 1, 2 and 3, n individuals are chosen at random and asked a "yes-or-no-question" in order to investigate the attitudes of the social classes to that question. The results could be summarized as follows:

Observed frequencies

	Social class		
Answer	1	2	3
Yes	n_{11}	n_{12}	n_{13}
No	n_{21}	n_{22}	n_{23}
Sums	n	n	n

[1] For convenience it is preferable to use the term nominal scale even though it may not be correct to call it a scale of measurement at all.

It is obvious how the χ^2-test will be made. However, if three interviewers, one from each social class, directed a question that is socially rather delicate to one and the same group of n individuals, we would have three dependent samples, and the above χ^2-test would not be applicable. Cochran (3) has in 1950 published a method, the Q-test, that is a modification of the χ^2-test, and which may be used when the samples are no longer independent.

In psycho-physical applications, for example, it is often feasible to let each individual answer the same "question" several times. The simple development that is given below, in order to allow Cochran's test to cover the case with more than one observation per cell, does not seem to have been discussed in the literature on the subject, at least not in recent text books. Perhaps this is because the F-test is thought to be a good enough approximation in practice.

2.2.2. *An application (9)*

In an experiment b individuals had to judge whether they perceived the "onset time" for 8 different tones, generated by the combination of 2 frequencies, 2 durations and 2 intensities, as longer, 1, or shorter, 0, than for a certain standard tone. This standard tone was one of the 8, so one comparison was the standard with itself. It may be of some importance whether the individuals hear the standard tone or the "experimental" tone first, and therefore it was decided to make at least two replications, in half the number of which the standard tone was to come first.

If we assume that the individuals made r judgements for each of a tones, the results could be summarized as in table IV: 1. The x's stand for the number of 1's. Thus individual no. j has judged x_{ij} times the "onset time" of the ith tone as longer than that of the standard tone. The question we want answered is then: is the perception of the "onset time" different for different tones? We want to find a test method applicable when $r=1$, and thus all x_{ij}'s either 0 or 1 (unrealistic in this application but relevant in others), and when r is a small number >1 and thus x_{ij} an integer between 0 and r.

2.2.3. *Cochran's Q-test and an extension to it*

When $r=1$, Cochran's Q-test may be applied to the problem in the preceding paragraph. In deriving the formula Cochran starts by building up an exact permutation test. The null hypothesis states that for any individual the probability of the answer "1" is constant, independent of tone. Of course it is possible that this probability varies from individual

TABLE IV: 1

Experimental tone no.	Individual no. 1	2	...	j	...	b	Sums
1	x_{11}	x_{12}	...	x_{1j}	...	x_{1b}	$x_{1.}$
2	x_{21}	...	...	...	...	...	$x_{2.}$
⋮							⋮
i	x_{i1}	...	...	x_{ij}	...	...	$x_{i.}$
⋮							⋮
a	x_{a1}	...	...	...	...	x_{ab}	$x_{a.}$
Sums	$x_{.1}$	...	...	$x_{.j}$	...	$x_{.b}$	$x_{..}$

to individual. Given the column[1] sums $x_{.j}$, that is, the number of 1's per individual, and the null hypothesis, the observed 1's could have been distributed over the rows in any (equally probable) way. In this way $\binom{a}{x_{.j}}$ possible permutations of column j is obtained, and for the whole experiment $\pi_j\binom{a}{x_{.j}}$ permutations, many of which, however, will coincide. It is then theoretically possible to compute $\Sigma_i(x_{i.}-(1/a)\Sigma_i x_{i.})^2$ for every permutation, and to find out where in this distribution the sum of squares for the observed permutation has fallen.

Cochran thereafter derives the limit distribution when the number of columns, b, increases. Taken for all the possible permutations the following is true:

$$E(x_{ij})=\frac{x_{.j}}{a};\quad D^2(x_{ij})=\frac{x_{.j}}{a}\left(1-\frac{x_{.j}}{a}\right);\mathrm{Cov}\,(x_{ij}\,;x_{i'j})=\frac{-x_{.j}/a(1-x_{.j}/a)}{a-1};\quad i\neq i'$$

The last equality follows from symmetry, the covariances must be equal for any pair of rows, and from the fact that the column sums $\Sigma_i x_{ij}=x_{.j}$ are considered constant.

If there is a large number of columns having the probability of giving the answer "1" not too near 0 or 1, the central limit theorem may be applied and thus the $x_{i.}$'s are approximately normally distributed. Further they obey the following:

$$E(x_{i.})=\sum_j\frac{x_{.j}}{a};\quad D^2(x_{i.})=\sum_j\frac{x_{.j}}{a}\left(1-\frac{x_{.j}}{a}\right);\quad \mathrm{Cov}(x_{i.}\,;x_{i'.})=-\frac{1}{a-1}\sum_j\frac{x_{.j}}{a}\left(1-\frac{x_{.j}}{a}\right)$$

Finally he utilizes the fact that, if the $x_{i.}$'s, $i=1, 2 \ldots a$, have a multivariate normal distribution with common variance σ^2 and covariance $\varrho\sigma^2$,

[1] The test is constructed so that for every given choice of column sums, the conditional probability of rejecting a true hypothesis is $\leqslant\alpha$. Then the unconditional risk of committing an error of the first kind must be $\leqslant\alpha$ too.

then: $\Sigma_i(x_{i.}-(1/a)\Sigma_i x_{i.})^2$ is distributed as: $\chi^2\cdot\sigma^2(1-\varrho)$, with $a-1$ degrees of freedom (19).

Here we have $\sigma^2(1-\varrho)=1/(a-1)\sum_j x_{.j}(1-x_{.j}/a)$, and thus the test variable:

$$Q=\frac{(a-1)\sum_i(x_{i.}-1/a\sum_i x_{i.})^2}{\sum_j x_{.j}(1-x_{.j}/a)}=\frac{(a-1)[a\sum_i x_{i.}^2-(\sum_i x_{i.})^2]}{a\sum_j x_{.j}-\sum_j x_{.j}^2}$$

When $r>1$ we obtain

$$E(x_{ij})=r\cdot\frac{x_{.j}}{ra};\quad D^2(x_{ij})=\frac{ra-r}{ra-1}\cdot r\cdot\frac{x_{.j}}{ra}\left(1-\frac{x_{.j}}{ra}\right)$$

$$\mathrm{Cov}(x_{ij};x_{i'j})=-\frac{1}{a-1}\cdot\frac{ra-r}{ra-1}\cdot r\cdot\frac{x_{.j}}{ra}\left(1-\frac{x_{.j}}{ra}\right)$$

(Note that the finite population correction disappears when $r=1$.)

This gives

$$\sigma^2(1-\varrho)=\frac{1}{a-1}\cdot\frac{ra-r}{ra-1}\sum_j x_{.j}\left(1-\frac{x_{.j}}{ra}\right)$$

and

$$Q=\frac{(ra-1)[a\sum_i x_{i.}^2-(\sum_i x_{i.})^2]}{ra\sum_j x_{.j}-\sum_j x_{.j}^2}$$

which is distributed approximately as χ^2 with $a-1$ degrees of freedom.

2.2.4. *Another derivation of the test formula.*

It is also possible to use another derivation of the Q-formula. This derivation, however, does not point out as clearly as the first one the possibility of checking the accuracy of the χ^2-test via an exact permutation test.

We assume that there is a certain probability for every individual of saying, for every experimental tone, that he perceives its "onset time" as longer than that of the standard tone. These probabilities are shown in table IV: 2.

The null hypothesis that the tones do not differ as regards the perception of "onset time" can be written:

$$H_0\colon p_{1j}=p_{2j}=\ldots=p_{ij}=\ldots=p_{aj}=p_j,\quad \text{for all } j\text{'s} \qquad \text{(IV.1)}$$

Observe that this hypothesis requires that the probabilities are equal within *every* individual. Alternatively the null hypothesis could express that the probabilities of the tones are equal *on the average* for the individuals in some population of individuals, from which we have a random sample.

TABLE IV: 2

Experimental tone no.	Individual no. 1	2	...	j	...	b
1	p_{11}	p_{12}	...	p_{1j}	...	p_{1b}
2	p_{21}	...	...	...	...	...
⋮						
i	p_{i1}	...	...	p_{ij}	...	p_{ib}
⋮						
a	p_{a1}	...	...	...	...	p_{ab}

When the first mentioned null hypothesis is true, we could use the following model to explain how the various 0's and 1's have arisen: For individual no j we have an urn with infinitely many balls. On the proportion p_j of the balls is printed 1, on the rest 0. From this urn r balls are drawn for tone no. 1, r for tone no. 2 etc. In this way we obtain the integers $x_{1j} \dots x_{aj}$ that are mutually independent, and for which the relationships:

$$E(x_{ij}) = rp_j;\ D^2(x_{ij}) = rp_j(1-p_j)$$

hold.

We then consider the sums of x_{ij} for the different tones and denote them by $x_{i.} = \Sigma_j x_{ij}$. If the p_j-values do not come too close to 0 or 1, and b is large enough, we can apply the central limit theorem to the $x_{i.}$'s. As the individuals are independent, the $x_{i.}$'s are approximately normally distributed and obey the following: $E(x_{i.}) = r\Sigma_j p_j$ and $D^2(x_{i.}) = r\Sigma_j p_j(1-p_j)$. As x_{ij} and $x_{i'j}$ for $i \neq i'$ are uncorrelated, so are $x_{i.}$ and $x_{i'.}$.

Thus the variable:

$$\sum_{i=1}^{a} \frac{(x_{i.} - r\sum_j p_j)^2}{r\sum_j p_j(1-p_j)}$$

is distributed approximately as χ^2, with a degrees of freedom.

However, $\Sigma_j p_j$ and $\Sigma_j p_j(1-p_j)$ are not known but have to be estimated. Obviously p_j is estimated by $x_{.j}/ra$, where $x_{.j} = \Sigma_i x_{ij}$, and thus $r\Sigma_j p_j$ by $\Sigma_j(x_{.j}/a)$. The variance $\Sigma_j p_j(1-p_j)$ is estimated by

$$\frac{ra}{ra-1}\sum_j \frac{x_{.j}}{ra}\left(1 - \frac{x_{.j}}{ra}\right)$$

Therefore the test variable will be:

$$Q = \frac{\sum_i (x_{i.} - 1/a\sum_j x_{.j})^2}{r \cdot ra/(ra-1)\sum_j x_{.j}/ra(1 - x_{.j}/ra)} = \frac{(ra-1)\{a\sum_i x_{i.}^2 - (\sum_i x_{i.})^2\}}{ra\sum_j x_{.j} - \sum_j x_{.j}^2}$$

The common rule for determining the number of degrees of freedom in a χ^2-test is to subtract one for every independent parameter estimated. The two parameters, $\Sigma_j p_j$ and $\Sigma_j p_j(1-p_j)$, are, as a rule, not independently estimated. It is therefore natural to use $a-1$ degrees of freedom, which will give us exactly the same test as before. For some values of the p_j the two estimators $\Sigma_j x_{.j}$ and $\Sigma_j x_{.j}(ra-x_{.j})$ are completely independent, for others they are completely dependent or have a correlation <1. According to the rule the correct number of degrees of freedom would then vary between $a-1$ and $a-2$. To choose $a-1$ is to make the test conservative.

2.2.5. *The possibility of varying p-values*

When, as in the application discussed, there are several replications per tone and individual scattered over a rather large time interval, some "learning effects" may appear and change the p_{ij}-values, that is the probabilities of the answer "1". It is also possible that p_{ij} depends on whether the standard comes first or last. Varying p-values would invalidate the binomial assumptions used in the derivation of the Q-formula. However, the test will then become more conservative, as an average value of the probabilities will be used in practice, which will give too large a variance.

Let us introduce the notation

$$x_{ij} = \sum_{\nu=1}^{r} x_{ij\nu}$$

with $$E(x_{ij\nu}) = p_{ij\nu}$$

Then we obtain $$E(x_{ij}) = \sum_{\nu=1}^{r} p_{ij\nu} = r\bar{p}_{ij}$$

and $$D^2(x_{ij}) = \sum_{\nu} p_{ij\nu}(1-p_{ij\nu}) \leqslant r\bar{p}_{ij}(1-\bar{p}_{ij})$$

The last inequality can be obtained from

$$\sum_{\nu} p_\nu(1-p_\nu) - r\bar{p}(1-\bar{p}) = -\sum_{\nu} p_\nu^2 + r\bar{p}^2 = \frac{1}{r}(\sum_{\nu} p_\nu^2)^2 - \sum_{\nu} p_\nu^2 \leqslant 0$$

which is true because we have a second degree polynomial with a maximi-point $=0$ when all p_ν's are equal.

This result can also be applied to another variant of the test formula. If we consider the sums of 1's and 0's for all individuals, (see table on the next page) we can construct the test variable:

$$\chi^2 = \sum_i \frac{(x_{i.} - x_{..}/a)^2}{x_{..}/a} + \sum_i \frac{(x_{i.} - x_{..}/a)^2}{rb - x_{..}/a} = \frac{rb\sum_i (x_{i.} - 1/a\sum_i x_{i.})^2}{rb\sum_j x_{.j}/a - 1/a^2(\sum_j x_{.j})^2}$$

Tones	Sums of 1's	Sums of 0's	Totals
1	$x_{1.}$	$rb - x_{1.}$	$r\ b$
2	$x_{2.}$	$rb - x_{2.}$	$r\ b$
⋮	⋮	⋮	⋮
i	$x_{i.}$	$rb - x_{i.}$	$r\ b$
⋮	⋮	⋮	⋮
a	$x_{a.}$	$rb - x_{a.}$	$r\ b$
Totals	$x_{..}$	$rb - x_{..}$	$r\ b\ a$

Above we have estimated $b\bar{p}$ and $b\bar{p}(1-\bar{p})$ instead of $\Sigma_j p_j$ and $\Sigma_j p_j(1-p_j)$. Thus the variance is overestimated and the test conservative. As we have an ordinary contingency table, there must be $a-1$ degrees of freedom.

2.2.6. *Different hypotheses and models*

The null hypothesis used in the derivations of the test formula was that of equal probabilities within blocks (individuals). However, as can be seen from the formula, only the differences between tones on the average over blocks are tested, that is the hypothesis

$$H_0: \sum_j p_{1j} = \ldots = \sum_j p_{aj} \tag{IV.2}$$

With additivity we mean that

$$p_{i1} = p_{i2} = \ldots = p_{ib} \text{ for all } i,$$

that is, the probabilities for a "1" with tone i are equal for all blocks (individuals) in the population. When this is true it can be seen that the two above mentioned null hypotheses coincide.

When there is interaction the null hypotheses are different and then it is important to specify the model. Cochran's Q tests the hypothesis (IV.2) when the model is fixed, that is, when we do not try to make any inference beyond the blocks in the experiment. However, if the model is mixed, that is if the blocks in the experiment are considered as a random sample from a large population of blocks, the Q-test tends to overestimate the significance. The differences between rank sums are namely compared to the variances within cells. One can say that in a fixed model the Q-test reacts on differences in average effects, hypothesis (IV.2), while for a mixed model it even reacts on interaction, hypothesis (IV.1).

For large r-values it would be possible to use another test variable. One could compute a χ^2 for every individual in the following way:

Individual no.	Answer	Tone 1	2	...	a	Sums
j	0	$r-x_{1j}$	$r-x_{2j}$	...	$r-x_{aj}$	$ra-x_{\cdot j}$
	1	x_{1j}	x_{2j}	...	x_{aj}	$x_{\cdot j}$
Sums		r	r	...	r	ra

$$\chi^2=\sum_{i=1}^{a}\left[\frac{(x_{ij}-r\hat{p}_j)^2}{r\hat{p}_j}+\frac{\{r-x_{ij}-r(1-\hat{p}_j)\}^2}{r(1-\hat{p}_j)}\right]=\sum_{i=1}^{a}\frac{(x_{ij}-r\hat{p}_j)^2}{r\hat{p}_j(1-\hat{p}_j)}$$

and then add these to

$$\chi^2=\sum_{j=1}^{b}\sum_{i=1}^{a}\frac{(x_{ij}-r\hat{p}_j)^2}{r\hat{p}_j(1-\hat{p}_j)}$$

with $b(a-1)$ degrees of freedom.

This presupposes that the rp_j's are large; common requirements in text books are $rp_j\geqslant 5$ or $rp_j\geqslant 10$. Cochran (4) has, however, investigated the applicability of the χ^2-test using small expected values and has found that, if the number of classes is large, some of the expected values may be as small as 1 without invalidating the test. Even these requirements would probably be too strong for the application discussed above.

The last test formula includes $a-1$ degrees of freedom for average differences and $(b-1)(a-1)$ for interaction. Thus it reacts on both differences in average effects and on interaction even when the model is fixed. But if there is additivity and if the rp_j's are large, this test will be more sensitive than the earlier derived Q-test.

2.2.7. *The possibility of using F-tests*

Because of what has been said above about the trouble due to interaction, it would be valuable to be able to use F-tests instead of the discussed χ^2-tests. But using a nominal scale the variables are far from normally distributed, moreover, the variances will as a rule vary from cell to cell. It is therefore difficult to say if and when the F-test can be used. Cochran (3) has discussed the possibility in the case of additivity. Though he only considered the case when $r=1$, he obtained almost as good results using a continuity correction to the F-test as when χ^2 is used, if the results be compared to an exact permutation test.

When $r>1$, it is then natural to expect a still better approximation. However, it is possible that a deviation from additivity will have a nega-

TABLE IV: 3. *The number of times the experimental tone was judged longer than the standard tone. Experimental tone no. 1 is the same as the standard tone.*

	Tone no.							
Individual no.	1	2	3	4	5	6	7	8
A. *Onset time*								
1	3	5	2	3	4	6	5	4
2	2	4	4	5	2	0	3	4
3	6	5	6	4	2	0	0	2
4	1	5	5	6	0	3	4	4
5	3	3	5	6	3	3	6	6
6	2	1	5	6	1	0	6	6
7	2	2	1	3	4	3	2	2
8	4	6	4	4	2	5	2	5
9	4	4	6	5	0	0	3	2
10	5	2	2	2	5	2	4	3
B. *Decay time*								
1	3	4	6	5	1	0	0	1
2	4	4	4	3	0	0	0	0
3	4	5	5	2	1	2	0	1
4	4	4	3	1	2	1	2	1
5	3	3	3	3	3	0	1	1
6	3	2	4	2	3	0	3	2
7	2	1	3	0	2	2	2	1
8	3	3	3	1	2	1	1	1
9	4	0	6	1	0	0	0	0
10	2	6	4	3	2	2	2	2

tive effect. The analysis can be carried through either with the 0's and 1's as they stand, or with proportions of 1's. If the arc-sine transformation of the proportions is used approximate stability of variances is obtained. How these F-tests, with varying degrees of interaction and for different b's and r's, work will be studied with the help of simulation experiments in paragraph 2.2.9.

2.2.8. *A numerical example*

In the experiment both "onset time" and "decay time" were studied using 8 tones and 10 individuals. Six replications per tone and individual were made, completely randomized for each individual. In 3 of these the standard tone was heard first, in the other 3 the experimental tone. The results are shown in table IV: 3. To these data the three mentioned tests: the derived Q-test, the F-test on integers and the F-test on transformed proportions, have been applied. (See table IV: 4.)

The P-values are the probabilities of obtaining such large Q- or F-values just by accident. $P<0.10$, for example, means that the probability is less than 10% but greater than 5%.

TABLE IV: 4. *Results of the tests applied to the data in table IV: 3.*

A. χ^2-test: $Q = 29.68$; $P < 0.0005$

F-test on integers

Sources of variation	D.f.	MS	F	P
Individuals	9	0.47	2.47	< 0.01
Tones	7	1.03	2.06	< 0.10
Interaction	63	0.50	2.63	< 0.0005
Within	400	0.19		

F-test on transformed proportions

Sources of variation	D.f.	MS	F	P
Individuals	9	0.6061	3.64	< 0.0005
Tones	7	1.4535	2.20	< 0.05
Interaction	63	0.6609	3.96	< 0.0005
Within	∞	0.1667		

B. χ^2-test: $Q = 76.87$; $P < 0.0005$

F-test on integers

Sources of variation	D.f.	MS	F	P
Individuals	9	0.28	1.47	> 0.10
Tones	7	2.51	11.41	< 0.0005
Interaction	63	0.22	1.16	> 0.10
Within	400	0.19		

F-test on transformed proportions

Sources of variation	D.f.	MS	F	P
Individuals	9	0.4768	2.86	< 0.005
Tones	7	2.9571	9.93	< 0.0005
Interaction	63	0.3311	1.99	< 0.0005
Within	∞	0.1667		

The agreement between the two F-tests is not particularly good; perhaps the deviations from the assumptions are too large for at least one of them. However, the most interesting test is the one that involves tones, and the agreement is best in this case. The difference between the F-tests and the Q-test is still greater, at least in A. This can be partly explained by the fact that the F-tests are made as in the mixed model, which is the more realistic model in this case. Better agreement is obtained when F-tests are performed as in the fixed model.

A. χ^2: $Q/7 = 4.24$;[1] $P < 0.0005$;
F on integers: $F = 1.03/0.19 = 5.3$; $P < 0.0005$;
F on transformed proportions: $F = 1.4535/0.1667 = 8.7$; $P < 0.0005$.

[1] We divide by the number of degrees of freedom in order to facilitate the comparisons with the F-tests.

B. χ^2: $Q/7 = 10.98$; $P < 0.0005$;
F on integers: $F = 2.51/0.19 = 13.2$; $P < 0.0005$;
F on transformed proportions: $F = 2.9571/0.1667 = 17.7$; $P < 0.0005$.

In this experiment the standard tone was given 3 times before, and 3 times after the experimental tone. This gave the possibility of recording only the "majority figure" and of using Cochran's formula in its original form. For the sake of comparison the results are given below.

P-values

		1 observation/ cell	3 observations/ cell
A.	Standard before	> 0.50	< 0.10
	Standard after	< 0.30	< 0.005
B.	Standard before	< 0.001	< 0.0005
	Standard after	< 0.0005	< 0.0005

The significance is seen to be highly increased in A if all the observations are used.

The results may be disturbed by the fact that the individuals are not allowed to state that they perceive simply no difference.

2.2.9. *Some simulation experiments*[1]

In order to investigate the applicability of the usual F-test on observations made by using the nominal scale, some simulation experiments were made. Six different populations of the type discussed in paragraph 2.2.4 are constructed and their matrices of probabilities shown in table IV: 5. Tones and individuals have here been changed to the more general terms, treatments and blocks. The blocks are sampled with replacement, which makes the populations of blocks infinite, while all the treatments are used. Thus the model is mixed. For every sampled combination of block and treatment, r 1's and 0's are generated with the probability $p(IJ)$ for "1" occurring.

A population model can then be written as:

$$p(IJ) = p(\text{-}\,\text{-}) + \underbrace{\{p(I\text{-}) - p(\text{-}\,\text{-})\}}_{\text{Treatment effect}} + \underbrace{\{p(\text{-}J) - p(\text{-}\,\text{-})\}}_{\text{Block effect}}$$

$$+ \underbrace{\{p(IJ) - p(I\text{-}) - p(\text{-}J) + p(\text{-}\,\text{-})\}}_{\text{Interaction effect}}; \quad \begin{matrix} I = 1, 2 \ldots A \\ J = 1, 2 \ldots \infty \end{matrix}$$

[1] Performed on a CDC 3600. The program was constructed by Lars Bäckström.

TABLE IV: 5. *Probability matrices for six constructed populations 1–6. The probability $p(IJ)$ of treatment I and block J is the probability of the answer "1" in that cell.*

Block no.	Treatment no. 1	2	3	4	5
a) *Population 1*					
1	0.1	0.1	0.1	0.1	0.1
2	0.2	0.2	0.2	0.2	0.2
3	0.3	0.3	0.3	0.3	0.3
4	0.4	0.4	0.4	0.4	0.4
5	0.5	0.5	0.5	0.5	0.5
6	0.6	0.6	0.6	0.6	0.6
7	0.7	0.7	0.7	0.7	0.7
8	0.8	0.8	0.8	0.8	0.8
9	0.9	0.9	0.9	0.9	0.9
b) *Population 2*					
1	0.8	0.8	0.4	0.7	0.8
2	0.5	0.5	0.7	0.7	0.7
3	0.7	0.4	0.6	0.7	0.7
4	0.7	0.5	0.9	0.8	0.8
5	0.1	0.1	0.3	0.2	0.4
6	0.6	0.4	0.5	0.7	0.8
7	0.4	0.6	0.6	0.5	0.6
8	0.6	0.8	0.6	0.6	0.6
9	0.9	0.5	0.6	0.4	0.6
10	0.6	0.7	0.7	0.8	0.4
11	0.5	0.3	0.4	0.1	0.4
12	0.8	0.9	0.5	0.4	0.5
13	0.1	0.1	0.1	0.2	0.2
14	0.1	0.1	0.2	0.4	0.1
15	0.6	0.5	0.5	0.7	0.7
16	0.6	0.6	0.6	0.8	0.3
17	0.2	0.3	0.3	0.1	0.1
18	0.4	0.6	0.6	0.2	0.5
19	0.4	0.9	0.6	0.6	0.5
20	0.4	0.4	0.4	0.3	0.3
c) *Population 3*					
1	0.1	0.8	0.4	0.7	0.8
2	0.5	0.5	0.7	0.7	0.7
3	0.7	0.4	0.6	0.7	0.7
4	0.7	0.5	0.9	0.8	0.8
5	0.1	0.1	0.3	0.2	0.4
6	0.6	0.4	0.5	0.7	0.8
7	0.4	0.1	0.5	0.6	0.6
8	0.6	0.1	0.1	0.6	0.6
9	0.9	0.5	0.6	0.4	0.6
10	0.6	0.3	0.7	0.1	0.1
11	0.5	0.3	0.4	0.8	0.4
12	0.8	0.9	0.5	0.4	0.5
13	0.4	0.4	0.6	0.2	0.7
14	0.1	0.6	0.2	0.4	0.4
15	0.6	0.5	0.5	0.7	0.2

Table IV: 5 (*continued*).

Block no.	Treatment no. 1	2	3	4	5
16	0.6	0.6	0.6	0.8	0.3
17	0.2	0.7	0.3	0.6	0.3
18	0.8	0.6	0.6	0.2	0.5
19	0.4	0.9	0.6	0.3	0.5
20	0.4	0.8	0.4	0.1	0.1
d) *Population 4*					
1	0.9	0.9	0.9	0.1	0.1
2	0.1	0.1	0.1	0.9	0.9
3	0.8	0.8	0.8	0.2	0.2
4	0.2	0.2	0.2	0.8	0.8
5	0.7	0.7	0.7	0.3	0.3
6	0.3	0.3	0.3	0.7	0.7
7	0.9	0.9	0.8	0.2	0.1
8	0.1	0.1	0.2	0.8	0.9
9	0.6	0.6	0.6	0.4	0.4
10	0.4	0.4	0.4	0.6	0.6
11	0.9	0.8	0.7	0.3	0.2
12	0.1	0.2	0.3	0.7	0.8
13	0.8	0.7	0.6	0.2	0.1
14	0.2	0.3	0.4	0.8	0.9
15	0.7	0.6	0.5	0.3	0.2
16	0.3	0.4	0.5	0.7	0.8
17	0.6	0.5	0.4	0.3	0.2
18	0.4	0.5	0.6	0.7	0.8
19	0.5	0.4	0.3	0.7	0.8
20	0.5	0.6	0.7	0.3	0.2
e) *Population 5*					
1	0.1	0.5	0.8	0.0	0.9
2	0.9	0.1	0.5	1.0	0.3
3	0.0	0.9	0.1	0.5	0.7
4	1.0	0.3	0.9	0.1	0.5
5	0.5	0.7	0.2	0.9	0.1
f) *Population 6*					
1	0.1	0.1	0.1	0.9	0.9
2	0.9	0.9	0.9	0.1	0.1

Introducing technical errors, we have

$$y(IJ) = p(IJ) + e(IJ)$$

where the $e(IJ)$'s are uncorrelated random variables which take on values $1-p(IJ)$ and $-p(IJ)$, having probabilities of $p(IJ)$ and $1-p(IJ)$ respectively. Therefore:

$$E\{e(IJ)\} = 0$$

$$D^2\{e(IJ)\} = p(IJ)\{1-p(IJ)\}$$

The following definitions are also used:

$$\sigma_A^2 = \frac{1}{A-1} \sum_{I=1}^{A} \{p(I\text{-}) - p(\text{-}\text{-})\}^2$$

$$\sigma_B^2 = \frac{1}{B} \sum_{J=1}^{B} \{p(\text{-}J) - p(\text{-}\text{-})\}^2$$

$$\sigma_{AB}^2 = \frac{1}{(A-1)B} \sum_{I=1}^{A} \sum_{J=1}^{B} \{p(IJ) - p(I\text{-}) - p(\text{-}J) + p(\text{-}\text{-})\}^2$$

$$\sigma_e^2 = \frac{1}{AB} \sum_{I=1}^{A} \sum_{J=1}^{B} D^2\{e(IJ)\}$$

The reason for using $1/B$, instead of $1/(B-1)$, in σ_B^2 and σ_{AB}^2 is that the B different blocks, given in table IV: 5, all have the frequency $1/B$ in the infinite populations of blocks.

The six populations are constructed so that the hypothesis H_0: $\sigma_A^2 = 0$ is true. The classical assumptions, as discussed in chapter II.3, for using the F-test to examine this hypothesis, are that:

1) $e(IJ)$ are uncorrelated and normally distributed with constant variance;
2) the block effects $\{p(\text{-}J) - p(\text{-}\text{-})\}$, $J = 1, 2 \ldots$, constitute a normal population;
3) the covariance matrix $\sigma_{II'}$ (see II.3.2) is such that:

$$\sigma_{II} = \sigma^2$$

$$\sigma_{II'} = \varrho\sigma^2; \quad I \neq I'$$

The first two of these assumptions are not fulfilled for any of the six populations. The $e(IJ)$'s always have a two-point distribution and $D^2\{e(IJ)\}$ ranges between 0.09 and 0.25 for populations 1–4, between 0.00 and 0.25 for population 5, while it is always 0.09 for population 6. The distributions of the block effects are far from normal. Assumption 3) is fulfilled for population 1 only. The details are given in table IV: 6. From this table it can be seen that the interaction is increasing as the population number increases.

Empirical sampling distributions are derived for $b = 3$ and 8, $r = 1$, 2, 4, 6 and 8, and with a always $= 5 = A$. For each of these combinations the electronic computer has sampled 10,000 sets of blocks independently and computed F on integers $= F$; F on arc-sine transformed proportions $= F_t$; and Q. Their distributions are compared to exact F- and χ^2-distributions in table IV: 7.

a) *Population 1*

$$\sigma_A^2 = 0$$

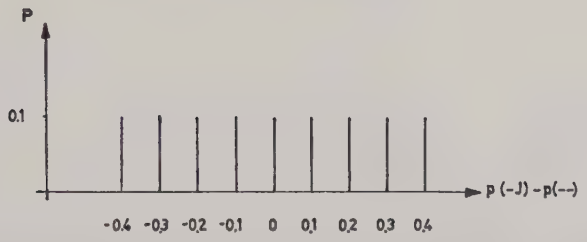
$$\sigma_{AB}^2 = 0$$

$$\sigma_B^2 = 0.33$$

$$\sigma_e^2 = 0.183$$

$$0.09 \leqslant D^2\{e(IJ)\} \leqslant 0.25$$

Distribution of block effects.



Covariance matrix $[\sigma_{II'}] = [0.33]$ (all terms in the matrix are $= 0.33$).

b) *Population 2*

$$\sigma_A^2 = 0$$

$$\sigma_{AB}^2 = 0.020$$

$$\sigma_B^2 = 0.0338$$

$$\sigma_e^2 = 0.2$$

$$0.09 \leqslant D^2\{e(IJ)\} \leqslant 0.25$$

Distribution of block effects.

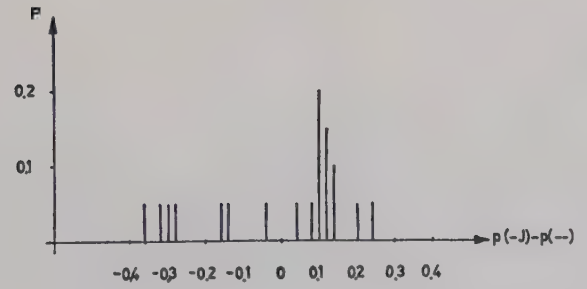


Table IV: 6 (*continued*).

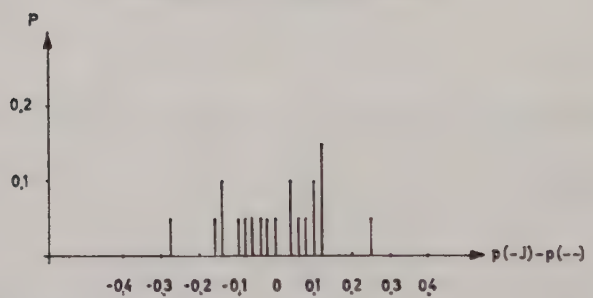
Covariance matrix $[\sigma_{II'}]$:

$$\begin{bmatrix} 0.058 & 0.037 & 0.028 & 0.032 & 0.036 \\ & 0.058 & 0.026 & 0.029 & 0.024 \\ & & 0.033 & 0.029 & 0.025 \\ & & & 0.058 & 0.033 \\ & & & & 0.047 \end{bmatrix}$$

c) *Population 3*

$$\sigma_A^2 = 0$$
$$\sigma_{AB}^2 = 0.044$$
$$\sigma_B^2 = 0.015$$
$$\sigma_e^2 = 0.20$$
$$0.09 \leqslant D^2\{e(IJ)\} \leqslant 0.25$$

Distribution of block effects.



Covariance matrix $[\sigma_{II'}]$:

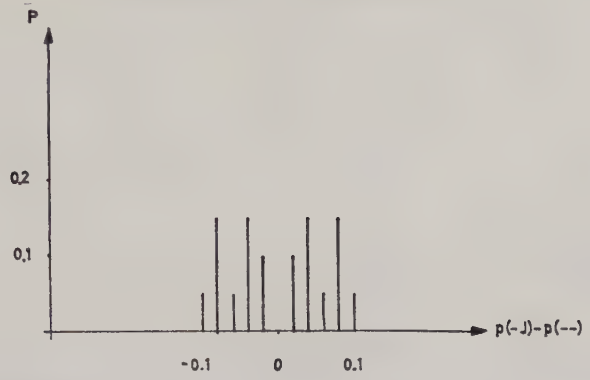
$$\begin{bmatrix} 0.054 & -0.001 & 0.023 & 0.006 & 0.006 \\ & 0.058 & 0.007 & -0.006 & -0.006 \\ & & 0.037 & 0.003 & 0.009 \\ & & & 0.058 & 0.022 \\ & & & & 0.047 \end{bmatrix}$$

d) *Population 4*

$$\sigma_A^2 = 0$$
$$\sigma_{AB}^2 = 0.0822$$
$$\sigma_B^2 = 0.0192$$
$$\sigma_e^2 = 0.1804$$
$$0.09 \leqslant D^2\{e(IJ)\} \leqslant 0.25$$

Table IV: 6 (*continued*).

Distribution of block effects.



Covariance matrix $[\sigma_{II'}]$:

$$\begin{bmatrix} 0.076 & 0.066 & 0.052 & -0.065 & -0.079 \\ & 0.061 & 0.052 & -0.058 & -0.069 \\ & & 0.049 & -0.048 & -0.055 \\ & & & 0.064 & 0.078 \\ & & & & 0.098 \end{bmatrix}$$

e) *Population 5*

$$\sigma_A^2 = 0$$
$$\sigma_{AB}^2 = 0.1444$$
$$\sigma_B^2 = 0.0128$$
$$\sigma_e^2 = 0.1324$$
$$0 \leqslant D^2\{e(IJ)\} \leqslant 0.25$$

Distribution of block effects.

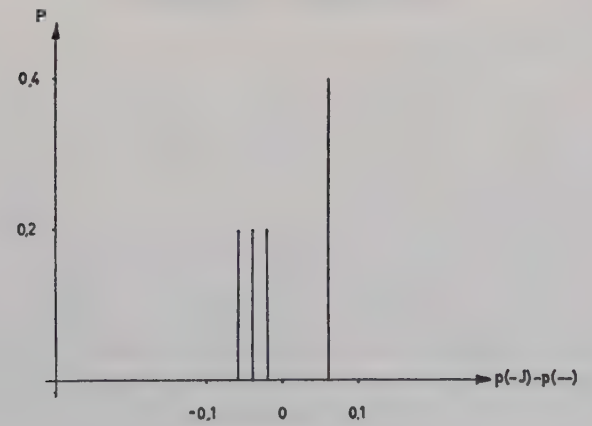


Table IV: 6 (*continued*).

Covariance matrix $[\sigma_{IT}]$:

$$\begin{bmatrix} 0.164 & -0.092 & 0.056 & 0.040 & -0.068 \\ & 0.080 & -0.060 & -0.008 & 0.016 \\ & & 0.100 & -0.086 & 0.032 \\ & & & 0.164 & -0.092 \\ & & & & 0.080 \end{bmatrix}$$

f) *Population 6*

$$\sigma_A^2 = 0$$

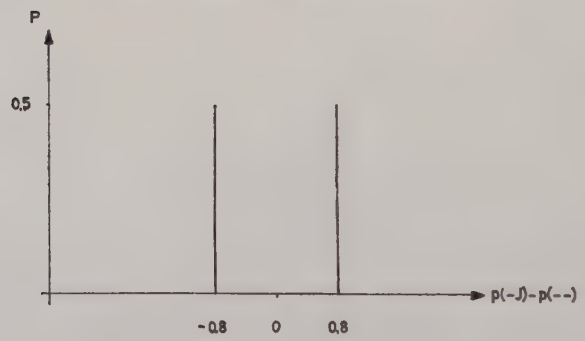
$$\sigma_{AB}^2 = 0.192$$

$$\sigma_B^2 = 0.0064$$

$$\sigma_e^2 = 0.09$$

$$D^2\{e(IJ)\} = 0.09$$

Distribution of block effects.



Covariance matrix $[\sigma_{IT}]$:

$$\begin{bmatrix} 0.16 & 0.16 & 0.16 & -0.16 & -0.16 \\ & 0.16 & 0.16 & -0.16 & -0.16 \\ & & 0.16 & -0.16 & -0.16 \\ & & & 0.16 & 0.16 \\ & & & & 0.16 \end{bmatrix}$$

Discussion of the results.—It is immediately seen that the results from population 6 strongly deviate from the other results. The main reason for this is probably that there are few values that are possible for F and F_t when there are only two different blocks in the population. A still more

"extreme" population can be obtained if the probabilities 0.1 and 0.9 in population 6 are changed to 0 and 1 respectively. It is then easy to compute the exact distribution for $F = MS_A/MS_{AB}$. For $b = 8$, $r = 8$, the following is obtained:

α	F
0.05	7.03
0.1	7.03
0.5	7.03
1.0	7.03
2.5	7.03
5	7.03
10	28.91
30	28.91

The distributions of F and F_t are very much alike. For $r = 1$ and 2 they are in fact identical, as the proportions 0, $\frac{1}{2}$ and 1 are transformed to 0, $\pi/2$ and π. When r becomes greater F_t seems to give a lower risk probability than F, especially for a low value of b. This means that F_t gives a somewhat better approximation than F for populations 2, 3, 4 and 6. For population 5, both F_t and F show overestimations when $b = 3$, and underestimations when $b = 8$. Thus F_t is the better test when $b = 3$, F the better when $b = 8$. In this sampling investigation F_t gives a definitely better approximation than F only in cases where both show a very bad approximation; otherwise they are equivalent. Thus it is difficult to say whether or not it is worth while to transform the proportions. However it seems to be somewhat safer to use F_t and therefore F will be neglected in the following study of the results.

In the more compact table IV: 8 the F_t-approximation can be studied for the most used levels of significance: 1%, 5% and 10%.

For populations 1–3, and, when $b = 8$, for populations 4 and 5, the approximation is very good. The results for population 6 when $b = 8$, indicate that if b is large, the F_t-test (or the F-test) can be used even for such a very extreme population if the test is interpreted with caution.

When b and r are small, the probability of observing $MS_{\text{Interaction}} = 0$ is not unimportant. Therefore rather many infinitely large F's and F_t's are obtained with ensuing overestimations as a result, at least for small α-values. Apart from this, increasing r's are in many cases accompanied by increasing observed frequencies of values above the significance points. In most cases we get overestimations (partly caused by the lack of symmetry discussed in II.3.2); this means that by increasing r we get a poorer approximation. Increasing b, on the other hand, gives a better approximation. The effect of the distribution of the block effects and the effect of increasing deviations from the above mentioned symmetry on the approximation

TABLE IV: 7. *Empirical distributions of F, F_t and Q, each based on 10,000 independent values. Observed frequencies (in %) outside upper α-limits for exact distributions are given.*

Nom. risk of error of type I $=\alpha\,\%$	r = number of observations (1's and 0's) per cell														
	1			2			4			6			8		
	F	F_t	Q	F	F_t	Q	F	F_t	Q	F	F_t	Q	F	F_t	Q
a) *Population 1*															
(i) $b=3$ blocks sampled with replacement															
0.05	1.90	1.90	1.81	0.21	0.21	0.10	0.01	0.07	0.02	0.06	0.04	0.05	0.03	0.04	0.04
0.1	1.90	1.90	1.81	0.27	0.27	0.10	0.04	0.09	0.02	0.10	0.12	0.08	0.04	0.08	0.06
0.5	2.16	2.16	1.81	0.70	0.70	0.12	0.46	0.50	0.25	0.56	0.57	0.32	0.34	0.46	0.42
1.0	2.40	2.40	1.81	0.90	0.90	0.26	0.87	0.93	0.65	1.01	1.14	0.71	0.73	0.89	0.96
2.5	2.56	2.56	1.90	1.91	1.91	1.20	2.20	2.34	1.91	2.30	2.50	2.06	2.17	2.18	2.47
5	6.52	6.52	2.40	4.61	4.61	3.15	4.33	4.56	4.13	4.53	4.72	4.65	4.67	4.69	4.83
10	8.78	8.78	6.52	8.79	8.79	8.59	9.01	9.05	8.85	9.59	9.71	9.67	10.06	10.24	10.06
30	26.35	26.35	31.36	29.59	29.59	30.79	29.83	29.54	29.97	29.56	29.64	30.15	30.90	30.51	31.04
(ii) $b=8$ blocks sampled with replacement															
0.05	0.04	0.04	0.00	0.06	0.06	0.01	0.10	0.05	0.02	0.05	0.03	0.06	0.04	0.01	0.03
0.1	0.09	0.09	0.00	0.10	0.10	0.03	0.13	0.12	0.08	0.08	0.08	0.10	0.07	0.05	0.04
0.5	0.47	0.47	0.09	0.37	0.37	0.24	0.54	0.57	0.48	0.38	0.46	0.43	0.51	0.58	0.46
1.0	1.02	1.02	0.33	0.78	0.78	0.69	1.00	1.16	0.98	0.84	0.94	0.83	1.01	1.04	0.88
2.5	2.30	2.30	1.44	2.38	2.38	2.11	2.55	2.46	2.25	2.29	2.19	2.20	2.62	2.60	2.50
5	5.57	5.57	4.10	4.87	4.87	4.30	4.99	4.98	4.71	4.73	4.75	4.45	5.13	5.17	5.16
10	10.31	10.31	8.86	10.07	10.07	9.67	9.56	9.77	9.62	9.66	9.81	9.56	10.08	10.38	10.37
30	30.11	30.11	30.43	30.25	30.25	30.38	30.16	30.22	29.62	29.87	29.79	29.67	30.29	30.14	30.33
b) *Population 2*															
(i) $b=3$ blocks sampled with replacement															
0.05	0.48	0.48	0.43	0.05	0.05	0.01	0.09	0.06	0.15	0.09	0.05	0.61	0.07	0.06	1.28
0.1	0.48	0.48	0.43	0.10	0.10	0.04	0.13	0.08	0.34	0.13	0.12	1.02	0.14	0.12	2.08
0.5	1.23	1.23	0.43	0.61	0.61	0.17	0.55	0.43	1.58	0.53	0.47	3.48	0.59	0.51	5.45
1.0	1.62	1.62	0.43	0.92	0.92	0.71	1.03	1.03	3.11	0.94	0.95	5.65	1.14	1.08	8.89
2.5	2.12	2.12	0.48	2.41	2.41	2.21	2.69	2.67	6.66	2.78	2.46	10.61	2.97	2.73	15.28
5	6.54	6.54	1.62	5.05	5.05	5.50	5.33	5.22	11.82	5.21	4.87	16.60	5.56	5.24	22.41
10	10.33	10.33	6.54	10.04	10.04	11.97	10.47	10.27	19.76	10.30	9.91	26.51	10.29	10.13	32.06
30	28.97	28.97	33.24	30.86	30.86	36.27	30.36	30.17	45.34	30.32	30.50	51.93	29.50	29.48	57.13

(ii) $b = 8$ blocks sampled with replacement

0.05	0.05	0.05	0.00	0.07	0.07	0.03	0.05	0.05	0.43	0.07	0.07	0.84	0.05	0.06	1.85
0.1	0.13	0.13	0.00	0.10	0.10	0.10	0.09	0.10	0.58	0.09	0.09	1.30	0.08	0.11	2.64
0.5	0.67	0.67	0.16	0.54	0.54	0.57	0.57	0.53	1.97	0.46	0.37	3.67	0.54	0.56	6.59
1.0	1.26	1.26	0.46	1.07	1.07	1.35	1.11	1.08	3.39	1.04	0.98	5.74	1.12	1.14	9.80
2.5	2.55	2.55	1.76	2.56	2.56	3.58	2.74	2.72	6.85	2.45	2.56	10.69	2.91	2.97	16.26
5	5.29	5.29	3.83	5.01	5.01	6.68	5.39	5.30	12.02	5.13	4.86	16.52	5.68	5.56	23.19
10	9.79	9.79	8.73	9.78	9.78	12.64	10.52	10.13	20.01	9.90	10.01	26.38	10.95	10.54	33.51
30	30.30	30.30	30.93	29.67	29.67	35.65	30.43	30.29	43.50	29.70	30.47	50.87	31.14	30.83	58.32

c) *Population 3*

(i) $b = 3$ blocks sampled with replacement

0.05	0.09	0.09	0.02	0.06	0.06	0.00	0.11	0.06	0.77	0.12	0.08	2.92	0.10	0.10	6.63
0.1	0.09	0.09	0.02	0.10	0.10	0.02	0.18	0.13	1.38	0.19	0.12	4.50	0.23	0.14	8.72
0.5	1.00	1.00	0.02	0.70	0.70	0.45	0.89	0.69	4.22	0.75	0.65	9.82	0.86	0.65	17.06
1.0	1.47	1.47	0.02	1.12	1.12	1.30	1.43	1.20	6.82	1.53	1.36	14.06	1.52	1.18	22.56
2.5	2.01	2.01	0.09	2.91	2.91	3.62	3.38	2.91	12.81	3.63	3.07	21.88	3.36	2.76	32.53
5	7.07	7.07	1.47	5.41	5.41	8.30	5.82	5.52	19.74	6.34	5.58	30.18	6.17	5.43	41.88
10	11.40	11.40	7.07	11.05	11.05	16.27	10.85	10.51	29.61	11.26	10.63	41.18	11.53	10.68	52.64
30	30.39	30.39	33.82	31.01	31.01	41.87	29.89	29.77	54.92	30.48	30.11	65.80	30.57	30.83	73.66

(ii) $b = 8$ blocks sampled with replacement

0.05	0.11	0.11	0.00	0.06	0.06	0.15	0.07	0.07	1.30	0.09	0.09	4.01	0.11	0.11	7.48
0.1	0.13	0.13	0.01	0.13	0.13	0.30	0.12	0.12	2.03	0.22	0.17	5.41	0.23	0.22	9.62
0.5	0.61	0.61	0.16	0.66	0.66	1.19	0.68	0.64	4.99	0.75	0.69	10.95	0.75	0.70	17.33
1.0	1.12	1.12	0.47	1.22	1.22	2.33	1.26	1.27	7.78	1.38	1.24	14.70	1.38	1.28	22.30
2.5	2.70	2.70	1.59	2.84	2.84	5.42	2.93	3.10	13.30	3.02	2.89	22.12	3.13	2.95	30.83
5	5.30	5.30	4.06	5.64	5.64	9.73	5.60	5.63	19.82	6.01	5.62	30.12	5.61	5.64	39.51
10	9.69	9.69	8.81	10.62	10.62	17.15	10.57	10.43	30.03	10.72	10.74	40.60	10.79	10.55	49.85
30	28.56	28.56	29.54	30.48	30.48	40.13	29.39	29.59	54.77	29.82	29.81	64.53	29.69	29.98	71.46

d) *Population 4*

(i) $b = 3$ blocks sampled with replacement

0.05	0.47	0.47	0.00	0.88	0.88	0.66	1.69	0.78	8.26	2.60	0.96	16.70	3.31	1.37	21.76
0.1	0.47	0.47	0.00	1.64	1.64	1.33	2.52	1.19	10.11	3.68	1.63	18.59	4.49	2.09	24.13
0.5	3.15	3.15	0.00	3.45	3.45	4.09	5.66	3.81	15.88	7.88	4.61	24.37	8.91	5.62	30.21
1.0	3.60	3.60	0.00	4.77	4.77	6.51	7.72	5.68	18.99	10.55	7.24	27.79	11.40	8.34	33.71
2.5	4.58	4.58	0.47	7.87	7.87	10.59	11.71	9.48	24.97	14.37	11.51	33.31	15.53	12.98	39.82
5	10.31	10.31	3.60	11.35	11.35	15.77	15.20	13.45	30.44	17.93	15.95	38.57	18.70	16.83	45.63
10	14.52	14.52	10.31	16.60	16.60	23.04	19.67	18.59	37.85	21.80	20.55	46.01	22.09	20.99	52.47
30	30.55	30.55	33.36	31.45	31.45	43.34	31.07	31.88	57.53	30.89	31.37	64.31	29.91	30.71	69.23

Nom. risk of error of type I $=\alpha\%$	r = number of observations (1's and 0's) per cell														
	1			2			4			6			8		
	F	F_t	Q	F	F_t	Q	F	F_t	Q	F	F_t	Q	F	F_t	Q
(ii) $b=8$ blocks sampled with replacement															
0.05	0.01	0.01	0.01	0.01	0.01	1.53	0.17	0.04	8.58	0.30	0.11	15.46	0.29	0.13	21.64
0.1	0.01	0.01	0.09	0.03	0.03	2.22	0.29	0.18	10.14	0.43	0.26	17.26	0.51	0.25	23.62
0.5	0.13	0.13	0.56	0.45	0.45	5.03	1.02	0.78	14.96	1.26	0.98	23.68	1.39	1.12	29.88
1.0	0.33	0.33	1.14	0.71	0.71	7.07	1.65	1.37	17.90	2.11	1.70	27.21	2.21	1.82	33.66
2.5	1.01	1.01	2.82	2.16	2.16	10.75	3.11	3.01	23.60	3.76	3.44	32.95	4.38	3.94	39.34
5	2.65	2.65	5.41	4.05	4.05	15.39	5.29	5.07	29.14	6.21	5.81	38.66	6.89	6.55	44.95
10	6.18	6.18	10.43	8.05	8.05	22.41	9.56	9.20	36.47	10.12	9.75	45.57	11.03	10.57	52.20
30	23.52	23.52	28.69	23.90	23.90	42.31	23.96	23.87	55.24	24.15	24.32	62.02	23.91	24.22	68.42
e) *Population 5*															
(i) $b=3$ blocks sampled with replacement															
0.05	0.37	0.37	0.00	0.79	0.79	0.40	1.75	1.44	15.28	2.87	2.48	29.75	3.25	3.06	38.64
0.1	0.37	0.37	0.00	1.29	1.29	1.48	2.45	2.10	18.33	3.16	3.02	32.53	3.63	3.48	42.88
0.5	2.86	2.86	0.00	2.98	2.98	5.00	4.02	3.82	26.92	4.45	4.23	42.26	4.61	4.49	54.01
1.0	3.12	3.12	0.00	4.16	4.16	9.55	5.27	4.92	31.42	5.34	5.17	47.31	5.47	5.36	59.19
2.5	3.96	3.96	0.37	6.59	6.59	16.05	7.99	7.69	39.54	7.78	7.85	55.68	8.08	7.88	66.29
5	9.83	9.83	3.12	10.30	10.30	23.40	11.23	10.93	46.96	12.27	11.70	62.43	13.14	12.30	71.63
10	14.19	14.19	9.83	15.30	15.30	33.27	16.72	16.29	56.36	18.76	17.77	69.46	19.56	18.72	76.67
30	31.84	31.84	34.54	31.43	31.43	54.47	29.99	29.91	74.14	30.55	30.63	82.96	29.53	29.93	87.53
(ii) $b=8$ blocks sampled with replacement															
0.05	0.00	0.00	0.02	0.00	0.00	2.20	0.02	0.03	14.78	0.05	0.04	28.36	0.04	0.03	39.25
0.1	0.00	0.00	0.04	0.01	0.01	3.13	0.06	0.06	17.30	0.09	0.09	31.38	0.06	0.05	42.77
0.5	0.07	0.07	0.46	0.12	0.12	6.97	0.31	0.28	26.28	0.48	0.46	40.77	0.36	0.32	51.82
1.0	0.22	0.22	1.02	0.44	0.44	9.71	0.67	0.61	31.23	0.73	0.79	45.51	0.81	0.74	56.32
2.5	0.91	0.91	2.90	1.28	1.28	15.34	1.80	1.71	39.01	1.89	1.82	52.76	2.19	2.21	63.24
5	2.57	2.57	5.59	3.02	3.02	22.03	3.45	3.38	46.13	3.83	3.74	59.47	4.33	4.18	69.04
10	6.42	6.42	11.04	6.94	6.94	30.33	7.41	7.33	55.17	7.79	7.61	66.79	8.77	8.57	75.22
30	25.59	25.59	30.78	24.68	24.68	52.92	24.84	24.96	72.93	24.67	24.56	81.37	25.10	25.04	86.53

f) *Population 6*

(i) $b = 3$ blocks sampled with replacement

0.05	5.35	5.35	0.00	10.84	10.84	10.51	19.51	13.36	24.76	22.74	14.80	24.44	24.14	17.68	25.11
0.1	5.35	5.35	0.00	13.59	13.59	14.42	21.40	16.61	24.82	23.60	18.42	24.47	24.45	20.97	25.68
0.5	14.65	14.65	0.00	18.94	18.94	21.30	24.13	22.35	24.96	24.36	23.57	25.87	24.60	24.36	30.92
1.0	15.15	15.15	0.00	20.51	20.51	22.69	24.50	23.83	25.27	24.37	24.26	27.82	24.61	24.55	36.38
2.5	15.74	15.74	5.35	22.78	22.78	24.00	24.77	24.62	26.81	24.38	24.38	33.77	24.61	24.61	48.63
5	20.45	20.45	15.15	23.82	23.82	25.07	24.83	24.80	30.25	24.39	24.39	42.83	24.61	24.61	62.37
10	22.58	22.58	20.45	24.45	24.45	26.73	24.83	24.83	37.34	24.39	24.39	56.87	24.61	24.61	75.98
30	27.71	27.71	28.79	26.01	26.01	37.09	24.91	25.07	65.76	24.40	24.45	85.53	24.61	24.62	95.12

(ii) $b = 8$ blocks sampled with replacement

0.05	2.28	2.28	0.59	3.08	3.08	7.39	4.57	3.78	25.55	5.59	4.22	30.09	6.15	4.38	33.14
0.1	2.60	2.60	0.93	3.91	3.91	8.67	5.67	4.63	26.97	6.57	5.15	30.75	6.79	5.33	35.69
0.5	4.37	4.37	2.69	6.06	6.06	13.49	7.09	6.70	29.37	7.23	6.97	35.44	7.09	6.86	45.60
1.0	5.59	5.59	4.06	7.02	7.02	17.07	7.35	7.38	30.77	7.29	7.39	39.47	7.10	7.14	51.58
2.5	7.35	7.35	6.23	8.61	8.61	22.01	8.45	8.66	34.15	8.03	8.72	47.75	7.51	8.37	60.21
5	10.10	10.10	9.12	11.15	11.15	26.04	11.03	11.19	39.07	10.99	11.47	55.44	10.35	11.22	65.93
10	13.93	13.93	13.49	15.89	15.89	30.63	17.30	16.60	47.16	19.16	17.44	62.45	19.46	17.40	69.79
30	26.15	26.15	26.97	27.88	27.88	44.87	28.68	28.05	64.06	29.39	28.70	71.57	29.07	28.56	72.38

TABLE IV: 8. *Observed frequencies in % of F_i-values above the exact upper 1%-, 5%-, and 10%-limits.*

Population no.	b	α %	r = number of observations per cell				
			1	2	4	6	8
		1	2.40	0.90	0.93	1.14	0.89
	3	5	6.52	4.61	4.56	4.72	4.69
		10	8.78	8.79	9.05	9.71	10.24
1							
		1	1.02	0.78	1.76	0.94	1.04
	8	5	5.57	4.87	4.98	4.75	5.17
		10	10.31	10.07	9.77	9.81	10.38
		1	1.62	0.92	1.03	0.95	1.08
	3	5	6.54	5.05	5.22	4.87	5.24
		10	10.33	10.04	10.27	9.91	10.13
2							
		1	1.26	1.07	1.08	0.98	1.14
	8	5	5.29	5.01	5.30	4.86	5.56
		10	9.79	9.78	10.13	10.01	10.54
		1	1.47	1.12	1.20	1.36	1.18
	3	5	7.07	5.41	5.52	5.58	5.43
		10	11.40	11.05	10.51	10.63	10.68
3							
		1	1.12	1.22	1.27	1.24	1.28
	8	5	5.30	5.64	5.63	5.62	5.64
		10	9.69	10.62	10.43	10.74	10.55
		1	3.60	4.77	5.68	7.24	8.34
	3	5	10.31	11.35	13.45	15.95	16.83
		10	14.52	16.60	18.59	20.55	20.99
4							
		1	0.33	0.71	1.37	1.70	1.82
	8	5	2.65	4.05	5.07	5.81	6.55
		10	6.18	8.05	9.20	9.75	10.57
		1	3.12	4.16	4.92	5.17	5.36
	3	5	9.83	10.30	10.93	11.70	12.30
		10	14.19	15.30	16.29	17.77	18.72
5							
		1	0.22	0.44	0.61	0.79	0.74
	8	5	2.57	3.02	3.38	3.74	4.18
		10	6.42	6.94	7.33	7.61	8.57
		1	15.15	20.51	23.83	24.26	24.55
	3	5	20.45	23.82	24.80	24.39	24.61
		10	22.58	24.45	24.83	24.39	24.61
6							
		1	5.59	7.02	7.38	7.39	7.14
	8	5	10.10	11.15	11.19	11.47	11.22
		10	13.93	15.89	16.60	17.44	17.40

TABLE IV: 9. *Power in % for the Q-test of the hypothesis* H_0: $\sigma^2_{AB} = 0$ *when* $\sigma^2_A = 0$.

Population no.	σ^2_{AB}	b	α %	r = number of observations per cell: 1	2	4	6	8
			1	0.43	0.71	3.11	5.65	8.89
		3	5	1.62	5.50	11.82	16.60	22.41
			10	6.54	11.97	19.76	26.51	32.06
2	0.020							
			1	0.46	1.35	3.39	5.74	9.80
		8	5	3.83	6.68	12.02	16.52	23.19
			10	8.73	12.64	20.01	26.38	33.51
			1	0.02	1.30	6.82	14.06	22.56
		3	5	1.47	8.30	19.74	30.18	41.88
			10	7.07	16.27	29.61	41.18	52.64
3	0.044							
			1	0.47	2.33	7.78	14.70	22.30
		8	5	4.06	9.73	19.82	30.12	39.51
			10	8.81	17.15	30.03	40.60	49.85
			1	0.00	6.51	18.99	27.79	33.71
		3	5	3.60	15.77	30.44	38.57	45.63
			10	10.31	23.04	37.85	46.01	52.47
4	0.0822							
			1	1.14	7.07	17.90	27.21	33.66
		8	5	5.41	15.39	29.14	38.66	44.95
			10	10.43	22.41	36.47	45.57	52.20
			1	0.00	9.55	31.42	47.31	59.19
		3	5	3.12	23.40	46.96	62.43	71.63
			10	9.83	33.27	56.36	69.46	76.67
5	0.1444							
			1	1.02	9.71	31.23	45.51	56.32
		8	5	5.59	22.03	46.13	59.47	69.04
			10	11.04	30.33	55.17	66.79	75.22
			1	0.00	22.69	25.27	27.82	36.38
		3	5	15.15	25.07	30.25	42.83	62.37
			10	20.45	26.73	37.34	56.87	75.98
6	0.192							
			1	4.06	17.07	30.77	39.47	51.58
		8	5	9.12	26.04	39.07	55.44	65.93
			10	13.49	30.63	47.16	62.45	69.79

can be studied for populations 4–6. Only one value of a, the number of treatments, has been considered, but the approximation cannot be expected to be poorer for larger a-values.

The Q-test is a test of the hypothesis H_0: $\sigma^2_A = \sigma^2_{AB} = 0$. From population 1, where both main-effect and interaction are zero, one can see that Q gives underestimations for small values of $b \cdot r$, and a good approximation for large values. As a limit to an approximation that is deemed satisfactory

we can take something in between 10 and 20. The other five populations give indications of Q's power in testing the interaction. In table IV: 9, for the levels of significance 1%, 5% and 10%, the power is apparently increasing as r increases, but hardly at all as b increases (except for very small r's).

2.3. Summary

A particular example has been thoroughly discussed in this chapter, but the conclusions are, of course, applicable to any randomized-block-situation with one treatment factor when the scale of measurement is nominal. It can be of interest to study the following two hypotheses:

1) that the treatments are equal on every block;
2) that the treatments are equal on the average for all blocks in the population of blocks.

The first hypothesis includes zero interaction and could be tested by the extension to Cochran's Q-test[1] made above. It is wise to use this test only when the product of the number of blocks and the number of observations per cell is large, at least 10 to 20.

The other hypothesis just means that the main-effect is zero and can be tested by the usual F-test on arc-sine-transformed proportions. This test is accurate enough when the number of blocks is not too small. Probably as few as 5 to 10 blocks will give an excellent approximation, unless the distribution of block effects is extremely unnormal. In population 5 there were only four different values of the block effect in the population, and the approximation when $b=8$ was still acceptable. In population 6, using only two different values of the block effect the approximation was not good when $b=8$. However, the improvement over the case when $b=3$ indicates that the approximation would be acceptable for a somewhat larger b-value. Whether the population is distributed as poorly as population 6 or not, can probably be determined before the experiment.

[1] It is a test of the second hypothesis when the model is fixed.

CHAPTER 3

ORDINAL SCALE

3.1. Introduction

The measuring scale is said to be ordinal if it is possible to determine which of the relations $a>b$, $a=b$ and $a<b$ is true for two observations a and b, but not to give any absolute measure of the difference between them. Measurements of this kind are common in the behavioural sciences. A psychological test may give the correct rank order, but 10 points at one end of the scale may not mean the same as 10 points at the other end. When treating pure qualitative variables it is sometimes possible to rank "the observations", for example, to rank professions according to their social status. For such situations a series of rank tests have been worked out. In the following, some of them are discussed by considering hypotheses and models. In this discussion we do not always confine ourselves to the ordinal scale but study the methods even when the measurements are quantitative. The question of whether it is possible to use the ordinary parametric tests is also touched upon. In some situations it will be possible to apply, for example, F-tests to scores on a scale not much stronger than an ordinal scale or to ranks, which is equivalent. Of course, the hypothesis will then be nonparametric (qualitative).

3.2. Completely randomized designs with one factor

For this one-way situation the best test seems to be the Kruskal-Wallis H-test (10), which for two levels (groups) is essentially the same as the Mann-Whitney U-test (12). The last mentioned test is another version of Wilcoxon's test (20). The H-test uses the statistic:

$$H=\frac{\dfrac{12}{R(R+1)}\sum_{i=1}^{a}\dfrac{T_i^2}{r_i}-3(R+1)}{1-\sum_{\nu}(t_\nu^3-t_\nu)/(R^3-R)}$$

where T_i is the rank sum of the r_i observations at the ith level; $R=\Sigma_i r_i$ and t_ν the number of values in the νth group of ties. When the r_i's are

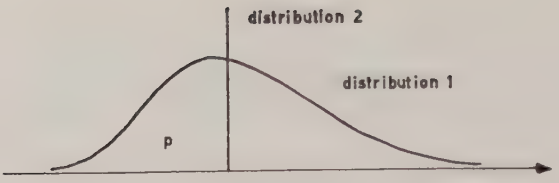


Fig. IV: 1.

large, H is approximately distributed as χ^2 with $a-1$ d.f.; for small r_i's there are tables of exact probabilities available in Mann–Whitney (12), Kruskal–Wallis (10) and Siegel (18).[1] If we consider observations from two distributions, x and y, ranked according to their algebraic magnitude, for example:

$$x\,x\,y\,x\,y\,y\,x\,y$$

the statistic U in the Mann–Whitney-test, counts the number of times an "x" is preceded by a "y". For the four x-values above we obtain:

$$U = 0+0+1+3 = 4$$

An important practical question is then: what is the null hypothesis in these tests? We are interested in knowing whether the a population means differ or not, and when some differ, which of them differ. In some text books the H- and U-tests are considered as tests for means and are sometimes recommended even when the measurements are quantitative, because they avoid most of the assumptions made in the parametric tests. But what do they avoid? This has been discussed by Kruskal–Wallis for instance. Theoretically the H-test involves a null hypothesis, which states that the a distributions are identical. The only classical assumption not used is that of normality, an assumption that is not very important unless the groups are very small. When the distributions are not identical in form the test of equal means (or equal medians) is affected. To get an idea of how much this test can be affected, let us consider the following extreme case:

Two populations are compared. One of them, 2, has zero variance, the other, 1, positive variance. This is the largest difference possible between two populations in this case (see fig. IV: 1). The proportion p of distribution 1 is located to the left of distribution 2.

If we want to use the U-statistic for testing the hypothesis of equal means

[1] When $a = 2$, up to groups of 8 observations, when $a = 3$, up to 5 observations per group.

as opposed to a double-sided alternative, the probability of committing a type-I-error is:

$$P(U \leqslant v \cdot r_2) = \sum_{x=0}^{v} \binom{r_1}{x} \{p^x (1-p)^{r_1-1} + p^{r_1-x} \cdot (1-p)^x\}$$

where r_1 and r_2 are the number of observations from distributions 1 and 2 respectively. The critical limit of U, as given in (18) is vr_2. The probability p is $=0.5$, if distribution 1 is symmetric, or if we are testing the hypothesis of equal medians.

$P(U \leqslant vr_2)$ has been computed for the nominal level of significance of 0.05, r_1 and $r_2 = 2(1)20$, and for $p = 0.5$ and 0.6. The critical values of U are taken from (18) and then v computed for the various values of r_1 and r_2. The results are shown in table IV: 10. The risks are seen to be large even for large values of both r_1 and r_2. The jumps in the tabulated probabilities are obtained when v is increased by 1. Knowing nothing about the variances the case where $r_1 = r_2$ is invariably chosen. In a test of equal medians there is an upper limit of the risk of a type-I-error of about 15%. This can also be considered as a limit of U's power when the alternative case is the one in which the variances are unequal. Knowing that the variance in one distribution is considerably less than in the other, one can take fewer observations from the distribution having a smaller variance and obtain a better agreement with the nominal 5%-level. Too few values can, however, give an extremely low power; see table IV: 10.

When the r_i's are large it is possible to use an ordinary F-test on the ranks. Fisher and Yates (7) have published a transformation of ranks in order to make them more suitable for the F-test. It is, however, difficult to see what would be gained by using the more laborious F-test, instead of the H-test in this simple model.

3.3. Randomized block designs with one factor

In this situation there is occasion to discuss main effects, treatment-block interaction and models. We shall see that the definitions of main effect and interaction are not exactly the same as in the quantitative case.

3.3.1. *Two levels*

The sign test.—The usual nonparametric test, when there are just two levels, is the wellknown sign test. This test is used when the observations are matched and only the signs for every pair of observations are recorded; i.e. when the difference between the two treatments is positive or negative (or zero). Using the hypothesis that positive and negative signs are equally

TABLE IV: 10. *The probability of wrongly rejecting the hypothesis of equal means (or when $p = 0.5$ the hypothesis of equal medians), using the U-test at the level 5 %, when distribution 2 has zero variance. Further r_1 observations are taken from distribution 1, r_2 from distribution 2. Also p is the probability of one observation from population 1 being less then the only possible value in population 2. When the same probability is obtained for several r_2-values, only the values corresponding to the smallest and largest r_2's are recorded.*

		r_2																
r_1	p	2	3	4	5	6	7	8	9	10	11	14	15	16	17	18	19	20
2	0.5	—	—	—	—	—	—	.50	—	—	—	—	—	—	—	—	—	.50
	0.6							.52	—	—	—	—	—	—	—	—	—	.52
3	0.5	—	—	—	.25	—	—	—	—	—	—	—	—	—	—	—	—	.25
	0.6				.28	—	—	—	—	—	—	—	—	—	—	—	—	.28
4	0.5	—	—	.125	—	—	—	—	—	—	—	—	—	—	—	—	—	.125
	0.6			.155	—	—	—	—	—	—	—	—	—	—	—	—	—	.155
5	0.5	—	.063	—	—	—	—	—	—	—	—	—	—	.063	.375	—	—	.375
	0.6		.088	—	—	—	—	—	—	—	—	—	—	.088	.424	—	—	.424
6	0.5	—	.031	—	—	—	.031	.219	—	—	—	—	—	—	—	—	—	.219
	0.6		.051	—	—	—	.051	.274	—	—	—	—	—	—	—	—	—	.274
7	0.5	—	.016	.016	.125	—	—	—	—	—	—	—	—	—	—	—	—	.125
	0.6		.030	.030	.177	—	—	—	—	—	—	—	—	—	—	—	—	.177
8	0.5	.008	.008	.070	—	—	—	—	—	—	—	—	—	.070	.289	—	—	.289
	0.6	.018	.018	.115	—	—	—	—	—	—	—	—	—	.115	.365	—	—	.365
9	0.5	.004	.004	.039	—	—	—	—	.039	.180	—	—	—	—	—	—	—	.180
	0.6	.010	.010	.074	—	—	—	—	.074	.257	—	—	—	—	—	—	—	.257
10	0.5	.002	.021	—	—	.021	.109	—	—	—	—	—	—	—	—	—	—	.109
	0.6	.006	.048	—	—	.048	.180	—	—	—	—	—	—	—	—	—	—	.180
11	0.5	.001	.012	—	.012	.065	—	—	—	—	—	—	—	.065	.227	—	—	.227
	0.6	.0037	.031	—	.031	.025	—	—	—	—	—	—	—	.125	.325	—	—	.325
12	0.5	.0005	.0063	.0063	.039	—	—	—	—	.039	.146	—	—	—	—	—	—	.146
	0.6	.0022	.020	.020	.087	—	—	—	—	.087	.241	—	—	—	—	—	—	.241
13	0.5	.0002	.0034	.0225	—	—	.0225	.092	—	—	—	—	—	—	—	—	—	.092
	0.6	.0013	.0128	.059	—	—	.059	.176	—	—	—	—	—	—	—	—	—	.176
14	0.5	.0001	.0018	.0129	—	.0129	.057	—	—	—	—	—	.057	.180[a]	.057[a]	.180	—	.180
	0.6	.0008	.0081	.041	—	.041	.129	—	—	—	—	—	.129	.297[a]	.129[a]	.297	—	.297
15	0.5	.0001	.0010	.0074	.0074	.035	—	—	—	.035	.118	—	—	—	—	—	—	.118
	0.6	.0005	.0052	.027	.027	.092	—	—	—	.092	.226	—	—	—	—	—	—	.226
16	0.5	.0000	.0042	.0042	.021	—	—	.021	.077	—	—	—	—	—	—	—	—	.077
	0.6	.0003	.018	.018	.066	—	—	.066	.172	—	—	—	—	—	—	—	—	.172
17	0.5	.0003	.0023	.0023	.013	.013	.049	—	—	—	—	.049	.143	—	—	—	—	.143
	0.6	.0021	.012	.012	.047	.047	.128	—	—	—	—	.128	.274	—	—	—	—	.274
18	0.5	.0001	.0013	.0075	.0075	.031	—	—	—	.031	.096	—	—	—	—	—	—	.096
	0.6	.0013	.0083	.033	.033	.096	—	—	—	.096	.215	—	—	—	—	—	—	.215
19	0.5	.0001	.0007	.0044	.0044	.019	—	.019	.064	—	—	—	—	—	—	—	—	.064
	0.6	.0008	.0055	.023	.023	.070	—	.070	.166	—	—	—	—	—	—	—	—	.166
20	0.5	.0000	.0004	.0026	.012	—	.012	.042	—	—	—	.042	.115	—	—	—	—	.115
	0.6	.0005	.0036	.016	.051	—	.051	.128	—	—	—	.128	.257	—	—	—	—	.257

[a] Probably some rounding-off effects in the critical values.

probable, it is possible to compute the probability of a sign distribution at least as disproportionate as that obtained. An early paper on this subject is Dixon–Mood 1946 (6).

We now consider a population consisting of B blocks. In each of these blocks the difference between the two factor levels has a certain distribution (possibly a one-point distribution). The sample consists of $b \leqslant B$ randomly chosen blocks for which a sign or zero is observed. The most natural model is a mixed one. We want to compare the average difference between the two factor levels in the experiment; the treatment factor is thus fixed. As a rule we want to generalize, i.e. make the results general for a larger population of blocks; that is $b < B$ and the block factor is random. When there is interaction it is a necessary condition that $b \ll B$ to be able to use the binominal distribution for the probability computations. When there is interaction in a fixed model, the sign test contains interaction as an "error term" and is thus conservative. An approximate test variable for this case will be proposed on pages 201–202.

The most common way of formulating the hypothesis, especially in more theoretical connections, is that the difference in every block has a continuous distribution and zero median. The probability is then $\frac{1}{2}$ for $+$, $\frac{1}{2}$ for $-$ and 0 for a tie. If the distributions are symmetric and identical the hypothesis is the same as in an ordinary t-test. If, in addition, the distributions are normal it is possible to compute the efficiency of the sign test compared to that of the t-test. However, the sign test is independent of the form of the distribution and does not depend on the constancy of the distribution from block to block. Another way of formulating the hypothesis is that there are as many plusses as minuses in the population. This can be the case when there is a certain rank order between the factor levels A_1 and A_2 for every block, and when $A_1 > A_2$ is true for half the number of blocks, while $A_1 < A_2$ for the rest. Then for every randomly chosen block: $P(+) = P(-) = \frac{1}{2}$. Note that we have introduced interaction between treatments and blocks; for some blocks the first treatment is best, for other blocks the second. This definition of interaction, suitable for the ordinal scale, is not exactly the same as the usual quantitative definition. Quantitative interaction may exist even if $A_1 > A_2$ for every block; namely when the quantitative difference varies in magnitude. The interaction defined for the ordinal scale corresponds to "nonremovable interaction"[1] in the quantitative case.

In practice it will sometimes happen that though there is a "true" sign

[1] By this is meant an interaction that can not be removed by any monotone transformation. See Scheffé (16) and Cox (5).

the wrong sign or zero is being observed. If the probability of wrong classification depends on how "deeply rooted" the signs are, or how great a quantitative value they correspond to, the null hypotheses will be moved in the quantitative direction.

Example 1. Let us say that the population consists of B blocks, each with a true difference D_J' between the two factor-levels. As we are concerned with the sign test we do not care about the numerical values of the differences, only about the signs. Therefore we introduce the differences D_J which are $+1$, 0 and -1 when the D_J''s are positive, zero and negative respectively. We also assume that the distributions of the differences are continuous so that the only possible values of D_J are ± 1. For every block there is a certain probability p_J of observing the wrong sign. To begin with we disregard the possibility of observing a tie.

Table over population values.

Block no.	1	2	...	J	...	B
"True" difference	D_1	D_2	...	D_J	...	D_B
Probability of observing the wrong sign	p_1	p_2	...	p_J	...	P_B

The blocks are arranged so that the first $B \cdot p$ D_J's are equal to $+1$, the remaining to -1. Thus p is the proportion of "plusses" in the population.

A simple random sample of b blocks, 1^x, 2^x ... j^x ... b^x is drawn and the differences d_1x, d_2x ... d_jx ... d_bx are observed. The sign test uses the probability of observing a plus (and a minus). We compute

$$P(d_jx = +1) = \sum_{J=1}^{Bp} P(d_jx = D_J)(1 - p_J) + \sum_{J=Bp+1}^{B} P(d_jx = D_J) \cdot p_J$$

$$= \frac{1}{B} \sum_{J=1}^{Bp} (1 - p_J) + \frac{1}{B} \sum_{J=Bp+1}^{B} p_J = p + \frac{1}{B} \left\{ \sum_{J=Bp+1}^{B} p_J - \sum_{J=1}^{Bp} p_J \right\}$$

When H_0 is true, $p = \frac{1}{2}$ and we can put $B = 2M$, thus:

$$P(d_jx = +1) = \tfrac{1}{2} + \frac{\sum_{J=1}^{M} (p_{M+J} - p_J)}{2M}$$

If, for example, the M positive differences have large numerical values and the M negative ones small numerical values, then it is reasonable to believe that: $\Sigma_{J=1}^{M}(p_{M+J} - p_J)$ is positive and $P(d_jx = +1) > \frac{1}{2}$. Thus there

will be a probability $>\alpha$ of rejecting the null hypothesis even when $p=\frac{1}{2}$. In the same way, if there are more positive signs than negative signs, but the positive signs are attached to smaller numerical values than the values attached to the negative signs, the probability of declaring $p>\frac{1}{2}$ will decrease.

If we allow the observation $d_j x=0$, we can use the following population table:

Block no.	1	2	...	J	...	B
"True" difference	D_1	D_2	...	D_J	...	D_B

For block J the probabilities of observing

$$\begin{Bmatrix} D_J \\ -D_J \\ 0 \end{Bmatrix} \text{ are: } \begin{Bmatrix} p_{1J} \\ p_{2J} \\ p_{3J} \end{Bmatrix} \text{ respectively}$$

From this it is easy to find:

$$P(d_j x=1 \,|\, d_j x \neq 0)=\frac{\sum_{J=1}^{M} p_{1J}+\sum_{J=M+1}^{2M} p_{2J}}{\sum_{J=1}^{2M}(p_{1J}+p_{2J})}$$

when H_0 is true.

Let us try to formulate a general hypothesis that covers all these cases. The only results that interest us are $+$, 0 and $-$. The distributions may therefore be characterized as three-point distributions. For block J there is a probability p_{1J} of observing $+$, p_{2J} of observing 0 and p_{3J} of observing $-$, where $p_{1J}+p_{2J}+p_{3J}=1$; $J=1, 2 \ldots B$.[1] (See fig. IV: 2).

Though this chapter has the title "Ordinal scale", we have discussed the sign test from a more general point of view than that of ordinal scales. We will now also go a little outside the scope indicated by the title of this paragraph ("Two levels").

The differences we are studying need not necessarily be of the simple type $x-y$, but can be arbitrary contrasts: $\Sigma_i c_i x_i$ with $\Sigma_i c_i=0$, for example: $2x-y-z$, or even linear functions of the type $2x-y$.

A natural definition of additivity is to say that blocks and treatments do not interact when $p_{1J}-p_{3J}$ is constant from block to block. It is easy to see that this definition covers the preceding cases.

[1] It is immaterial whether the possibilities of observing different signs are caused by observational errors or by some real fluctuations in time.

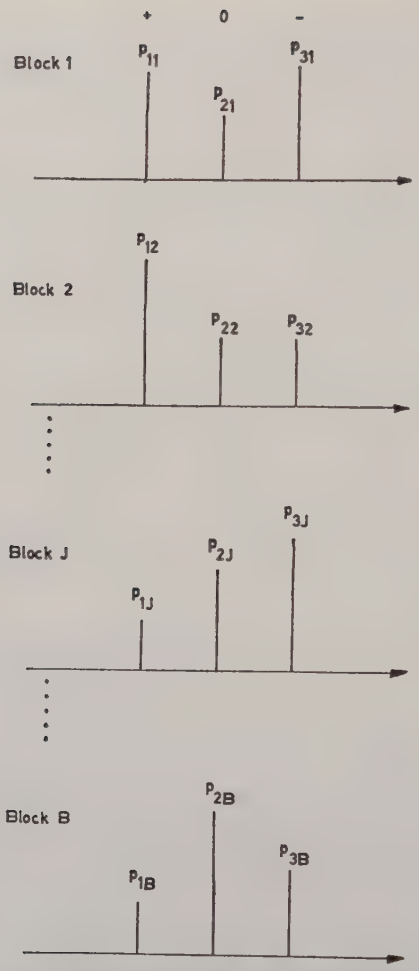


Fig. IV: 2. One three-point distribution of the "signs" +, 0, and − for each of the B blocks in the population.

In order to study the null hypothesis we first consider the case:

I. $$p_{2J} = 0; \quad J = 1, 2 \ldots B$$

The probability that a random observation will be + in block J, is

$$P(\text{Block } J \text{ and } +) = P(\text{Block } J) \cdot P(+ \mid \text{Block } J) = \frac{1}{B} \cdot p_{1J}$$

The probability of + is:

$$P(+) = \sum_J P(\text{Block } J \text{ and } +) = \frac{1}{B} \sum_{J=1}^{B} p_{1J}$$

and thus we test the hypothesis:

$$H_0: \sum_J p_{1J} = \sum_J p_{3J} \left(= \frac{B}{2} \right)$$

that is $P(+) = \frac{1}{2}$ when H_0 is true. We also have the following:

$$\sum_J p_{1J} > \sum_J p_{3J} \Rightarrow P(+) > \tfrac{1}{2}$$

$$\sum_J p_{1J} < \sum_J p_{3J} \Rightarrow P(+) < \tfrac{1}{2}$$

Thus the sign test will tend to give a significant result when the average median deviates from zero, but not when only the interaction is different from zero.[1]

In the following we will consider the more general case where it is possible that:

II. $\qquad p_{2J} > 0$ for some J's

Probably the most used strategy is then to ignore blocks with an observed zero. It is easy to compute:

$$P(+ \mid \text{not zero}) = \frac{\sum_J p_{1J}}{\sum_J p_{1J} + \sum_J p_{3J}}$$

that is, the hypothesis H_0: $\Sigma_J p_{1J} = \Sigma_J p_{3J}$ implies: $P(+ \mid \text{not zero}) = \frac{1}{2}$. It could be debated whether this is a reasonable hypothesis or not. As it can be written: $\Sigma_J(p_{1J} - p_{3J}) = 0$, one observes that, for example, the blocks J and J' having the following probabilities:

$$p_{1J} = 0.6 \quad p_{3J} = 0.4 \quad p_{1J} - p_{3J} = 0.2$$

$$p_{1J'} = 0.1 \quad p_{3J'} = 0.3 \quad p_{1J'} - p_{3J'} = -0.2$$

"cancel", though there is a smaller probability of obtaining block J' in the sample (blocks with a zero are ignored).

It is also apparent that large values of p_{2J} will decrease the power of the test.

Example 2. The following special population is considered:

---

[1] Sometimes one may be interested in testing the hypothesis that the average upper quartile (or any other fraction) is zero. The null hypothesis would then be formulated H_0: $3\sum_J p_{1J} = \sum_J p_{3J}$.

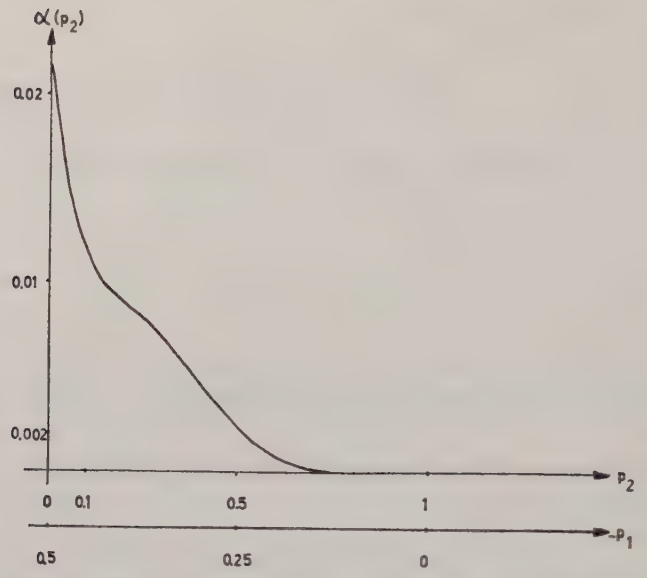


Fig. IV: 3. The risk of committing an error of type I, $\alpha(p_2)$, as a function of p_2 = the probability of a tie.

$$p_{1J} = p_1 \quad J = 1, 2 \ldots B$$
$$p_{2J} = p_2 \quad J = 1, 2 \ldots B$$
$$p_{3J} = p_3 \quad J = 1, 2 \ldots B$$

from which $b = 10$ blocks are sampled.

For a double-sided test at the level 2.5% the rejection region will consist of the following occurrences of signs:

+	0	−
10	0	0
9	0	1
9	1	0
8	2	0
7	3	0

and the corresponding region obtained by interchanging + and − above. The probability of rejecting the hypothesis that $p_1 = p_3$ when it is true, is denoted by $\alpha(p_2)$. It is shown as a function of p_2 in fig. IV: 3. It can be seen that already for p_2 as small as 0.1, α has dropped by almost 50%, which means that the power has also been considerably reduced.

When we are studying the contrast $2A_1 - A_2 - A_3$, and the observations are just ranks, there will exist a certain probability of observing ties even if we rank the three treatments without ties. The rank orders, 213 and 231,

will both give $2A_1-A_2-A_3$ a value of zero. If all 6 possible rank orders are equally probable the probability of a tie for the contrast is $\frac{1}{3}$. As in the example above, increasing probability of ties means decreasing power. For a nominal level of 2.5%, the true level can be computed or read from fig. IV: 3 as 0.6%. If the sign test is used for making multiple comparisons, and we are studying the contrast $2A_1-A_2-A_3$ when four treatments A_1, A_2, A_3 and A_4 have been ranked, we can either use the ranks between the three or between the four treatments. If there is an equal probability of observing all 24 possible rank orders, the probability of a tie is, in the latter case, reduced to $\frac{1}{6}$ with a better power as a result.

Let us now consider the case of a fixed model, i.e. where $b=B$. If there is interaction the sign test is then conservative. An alternative procedure can be devised as follows. Making n observations per block the matrix of observations can be denoted:

Block no.	Sign +	Sign 0	Sign −	Sums
1	x_{11}	x_{21}	x_{31}	n
2	x_{12}	x_{22}	x_{32}	n
⋮	⋮			⋮
j	x_{1j}	x_{2j}	x_{3j}	n
⋮	⋮			⋮
b	x_{1b}	x_{2b}	x_{3b}	n
Sums	x_1	x_2	x_3	bn

The probabilities are as on page 198. It is natural to use the difference x_1-x_3 and its expected value: $n\Sigma_{j=1}^{b}p_{1j}-n\Sigma_{j=1}^{b}p_{3j}=0$ when the null hypothesis is true. The variance for the difference is then:

$$D^2(x_1-x_3)=2\{D^2(x_1)+D^2(x_3)\}-D^2(x_2)$$

If it is estimated by:

$$2\left\{n\sum_j\frac{x_{1j}}{n}\left(1-\frac{x_{1j}}{n}\right)+n\sum_j\frac{x_{3j}}{n}\left(1-\frac{x_{3j}}{n}\right)\right\}-n\sum_j\frac{x_{2j}}{n}\left(1-\frac{x_{2j}}{n}\right)$$

we could use the test statistic:

$$\chi^2=\frac{(x_1-x_3)^2}{n\left[2\left\{\sum_j\frac{x_{1j}}{n}\left(1-\frac{x_{1j}}{n}\right)+\sum_j\frac{x_{3j}}{n}\left(1-\frac{x_{3j}}{n}\right)\right\}-\sum_j\frac{x_{2j}}{n}\left(1-\frac{x_{2j}}{n}\right)\right]}$$

which is approximately distributed as χ^2, with 1 d.f. when b is large. The statistic can, however, not be used when $n=1$. Moreover, the denominator is apt to be zero when n is small, even in cases where the difference x_1-x_3 is small. Therefore it is proposed to replace the variance:

$$D^2(x_i) = n\Sigma_j p_{ij}(1-p_{ij}), \text{ by } nb\bar{p}_i(1-\bar{p}_i)$$

which is an overestimate (see page 170). This gives the test variable:

$$\chi^2 = \frac{(x_1-x_3)^2}{bn\left[2\left\{\frac{x_1}{bn}\left(1-\frac{x_1}{bn}\right)+\frac{x_3}{bn}\left(1-\frac{x_3}{bn}\right)\right\}-\frac{x_2}{bn}\left(1-\frac{x_2}{bn}\right)\right]}$$

which is approximately distributed as χ^2, with 1 d.f. when b is large. The test is conservative but probably more sensitive than the sign test.

When the total distribution (obtained by "adding" all the distributions in all the blocks in the population) is symmetric, the above null hypothesis is the same as the usual quantitative hypothesis that the average mean is zero, otherwise it is not the same (compare the discussion of the Wilcoxon test below). When the distributions are continuous, symmetric, and not too different, and the number of blocks in the sample not too small, then the t-test may replace the sign test. The former has then greater power, but is somewhat more time-consuming.

Wilcoxon's test.—When it is possible to rank the differences one can use the Wilcoxon matched-pairs signed-ranks test (20, 18), which tests the main effect in the quantitative sense to a greater extent than does the sign test. The differences are ranked according to their "absolute values",[1] and then the ranks are given the same signs as the respective differences. The exact probability distribution of T, the smaller sum of ranks having the same sign, has been computed when a certain null hypothesis (to be discussed below) is true, and when b, the number of blocks, $\leqslant 25$. These sampling distributions can be found, for example, in (21).

By using the central limit theorem many text books state, that for large b's:

$$\lambda = \frac{T-b(b+1)}{\sqrt{\frac{b(b+1)\,(2b+1)}{24}}}$$

is approximately $N(0; 1)$. See, for example, (18).

[1] It is not necessary to be able to give numerical values to the differences, only to say that this difference, apart from the sign, is greater than that difference etc.

The null hypothesis for this test is often very vaguely formulated (or mathematically exact in a very restricted case, as has been discussed in connection with the sign test). Siegel (18) says in an example that the null hypothesis states that "the sum of positive ranks = the sum of negative ranks". In Brownlee's book (2), one can find the following: "Essentially the test is of the null hypothesis that the distribution of the differences is symmetric about zero, so any rank is equally likely to be positive or negative." In the computation of the exact distributions of T, the fact that any rank has the probability $\frac{1}{2}$ of being positive and the same probability of being negative is used.

Sometimes it is thought that Wilcoxon's test can be used as a test of the usual quantitative null hypothesis of zero average mean. To get an idea of what the hypothesis is, and the consequences of certain deviations, we will discuss some simple formal examples. The model will be the usual one, with B blocks in the population, and b in the random sample. In every block the difference has a certain distribution caused by time-fluctuations, observational errors etc. The qualitative null hypothesis can be defined in several ways, including the vague "as many plusses as minuses in the population". In case additivity occurs we use the definition "all medians zero". If there is interaction the qualitative null hypothesis could mean: "as many positive as negative means (or medians) in the population" or "average median zero". The last one is the one found to be tested by the sign test and will be used in the following arguments. Three situations, A–C, are discussed. In A there is no interaction between treatments and blocks, and for some examples the quantitative hypothesis is true, for others the qualitative. In B there is interaction and both types of hypotheses are true. There is also interaction in C, but only the quantitative null hypothesis is true in this case.

A. *No interaction between blocks and treatments.*

As there is some difficulty of defining additivity (no interaction) in the qualitative sense in situations for Wilcoxon's test, we have only considered the case where the blocks have identical distributions.

Example 3. Identical and symmetric distributions. (Fig. IV: 4).

If the hypothesis that the means are zero is tested against the hypothesis of a simple translation, then it is easy to realize that Wilcoxon's test works and has a better power than the sign test.

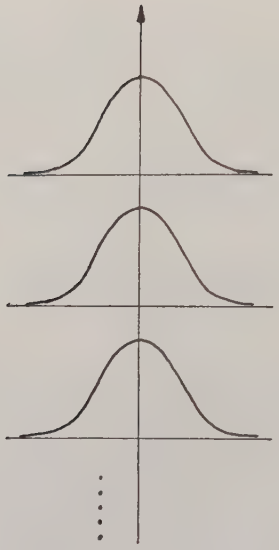


Fig. IV: 4.

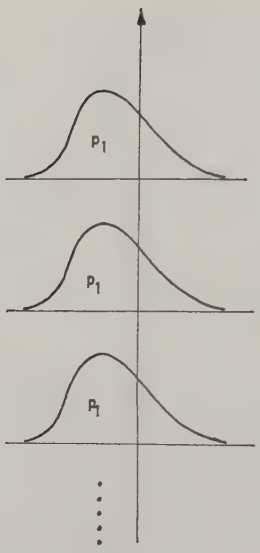


Fig. IV: 5.

Example 4. Identical but somewhat skew distributions. The means are zero. (Fig. IV: 5).

In the following examples we will compute the actual risks of a type-I-error, for some of the smallest values of b, for which significant results are possible. With a nominal level of 2.5% in a single-sided test, the critical value is 0 when $b=6$ (18).

If we denote in this example the probability of a negative value in each block by p_1, we get: $P(T_{\text{pos}}=0)=p_1^b$. With $b=6$ and $p_1=0.7$, this gives $P(T_{\text{pos}}=0)=11.76\%$ to be compared with the nominal 2.5%.

Example 5. Identical but somewhat skew distributions. The medians are zero. (Fig. IV: 6).

We consider here a distribution that consists of three continuous rectangular parts.

The one-sided test in the negative direction will have the largest distortion caused by the lack of symmetry and will therefore be investigated below. When $b=7$ a symmetric distribution gives:

$$P(T_{\text{pos}} \leqslant 2) = \frac{3}{2^7} = 0.0234$$

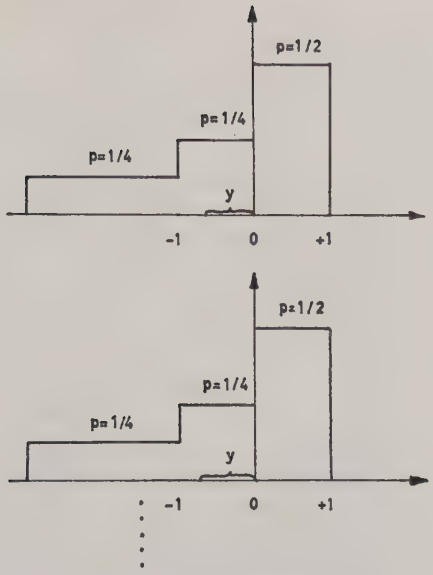


Fig. IV: 6.

With the distribution as in fig. IV: 6 we compute:

$$P(T_{\text{pos}}=0)=(\tfrac{1}{2})^b=0.0078 \quad \text{(all differences negative)},$$

$$P(T_{\text{pos}}=1)=\binom{b}{1}\cdot\left(\frac{1}{2}\right)^b\int_0^1\left\{\frac{1}{2}+\left(\frac{1-y}{2}\right)\right\}^{b-1}\cdot dy=\frac{2^b-1}{2^{2b-1}}=0.0155$$

(one difference positive and numerically smaller than all the negative differences),

$$P(T_{\text{pos}}=2)=\binom{b}{1}\left(\frac{1}{2}\right)^b\int_0^1(b-1)\left\{\frac{1}{2}+\left(\frac{1-y}{2}\right)\right\}^{b-2}\cdot\frac{y}{2}\,dy=\frac{2^b-1-b}{2^{2b-1}}=0.0146$$

(one difference positive and numerically smaller than all but one of the negative differences). Thus we obtain:

$$P(T_{\text{pos}}\leqslant 2)=0.0379$$

The distortions caused by moderate deviations from symmetry will probably not be large, at least not for small b's in double-sided tests.

Example 6. Identical bimodal distributions. The medians are zero. (Fig. IV: 7).

We consider the case where a negative difference is (practically) certain to have a smaller absolute value than a positive difference.

If one negative difference is observed, it will have the rank 1, with two their rank sum is 3, and so on.

The following comparisons between nominal and real significance levels can be easily computed:

b	Critical value	Real level for: One-sided test of positive median Nominal level ≤ 0.025	Real level for: Double-sided test Nominal level ≤ 0.05
6	0	$\left(\frac{1}{2}\right)^6 = 0.016$	$\left(\frac{1}{2}\right)^6 = 0.016$
7	2	$\left(\frac{1}{2}\right)^7 \sum_{j=0}^{1} \binom{7}{j} = 0.063$	$\left(\frac{1}{2}\right)^7 \left\{\sum_{j=0}^{1} \binom{7}{j} + \binom{7}{0}\right\} = 0.070$
8	4	$\left(\frac{1}{2}\right)^8 \sum_{j=0}^{2} \binom{8}{j} = 0.145$	$\left(\frac{1}{2}\right)^8 \left\{\sum_{j=0}^{2} \binom{8}{j} + \binom{8}{0}\right\} = 0.148$
9	6	$\left(\frac{1}{2}\right)^9 \sum_{j=0}^{3} \binom{9}{j} = 0.271$	$\left(\frac{1}{2}\right)^9 \left\{\sum_{j=0}^{3} \binom{9}{j} + \binom{9}{0}\right\} = 0.273$

This example gives an idea of the consequences when the greatest possible deviation from unimodal symmetric distributions occurs.

B. *Interaction. Both the quantitative and the qualitative null hypotheses are true.*

When we speak about identical distributions in the presence of interaction, we mean distributions identical in shape but having differences in location.

Example 7. Identical and symmetric distributions. (Fig. IV: 8).

Wilcoxon's test is exact as regards type-I-error and has a better power than the sign test.

Example 8. Different and somewhat skew distributions. (Fig. IV: 9).

The distributions are identical in shape except that half the number of blocks have positively skew distributions with negative medians and the other half negatively skew distributions with positive medians. The total distribution of all possible observable differences (obtained by projecting all distributions on the same horisontal axis) is symmetric with median

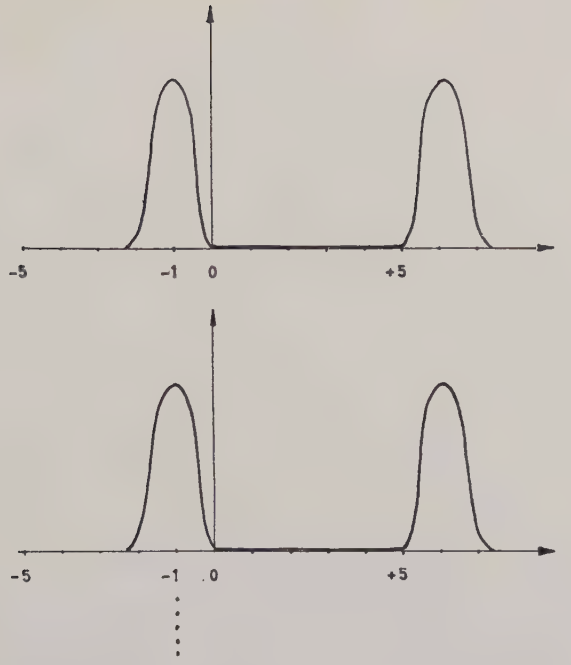


Fig. IV: 7.

zero. Thus all b ranks in the sample have an equal probability of a positive and a negative sign and the test is exact.

C. *Interaction. Only the quantitative null hypothesis is true.*

In examples 9 and 10 the population has as many blocks with negative as with positive means (medians).

Example 9. Identical, somewhat skew distributions. (Fig. IV: 10).

We immediately see that the probability of a negative difference is $>\frac{1}{2}$, which means that the sign test will tend to reject the null hypothesis in the negative direction; the qualitative null hypothesis is not true. For Wilcoxon's test we find

$$P(T_{\text{pos}}=0)=\sum_{\nu=0}^{b}\binom{b}{\nu}\left(\frac{1}{2}\right)^{\nu}p_1^{\nu}\cdot\left(\frac{1}{2}\right)^{b-\nu}\cdot p_2^{b-\nu}=\left(\frac{p_1+p_2}{2}\right)^{b}$$

$$\left.\begin{aligned}p_1&=0.85\\p_2&=0.35\\b&=6\end{aligned}\right\}\Rightarrow P(T_{\text{pos}}=0)=0.0467\quad\text{(nominal level 0.025)}$$

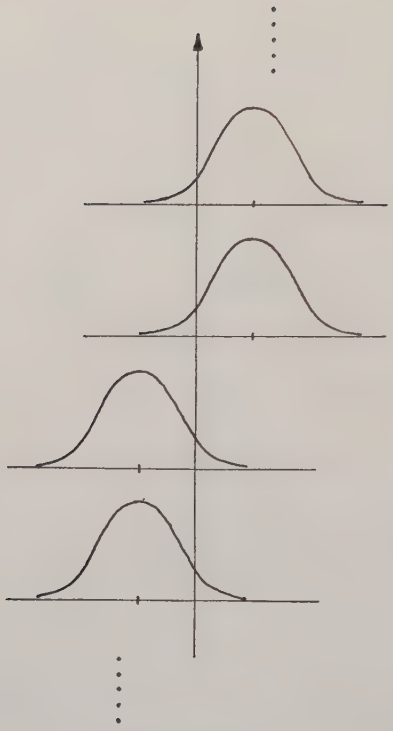


Fig. IV: 8.

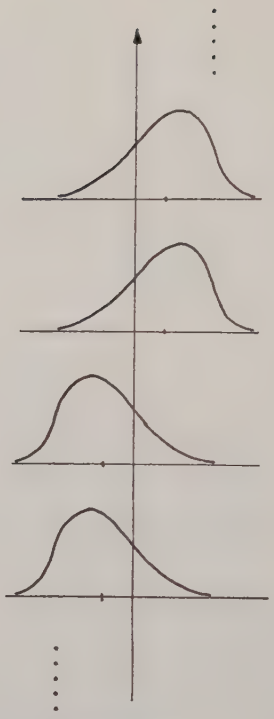


Fig. IV: 9.

In a two-sided test we get $P(T_{\text{neg}}=0)=\{(0.15+0.65)/2\}^6=0.0041$, and $P(T=0)=0.0508$, which is very near the nominal level 0.05. Apparently two-sided tests will not be affected as much as one-sided as regards type-I-errors. The power can, however, be decreased. For example it would be rather difficult to detect a small translation to the right in this example.

Example 10. Different distributions in different blocks.

When the distributions are not identical the test can be affected much more than in example 9. Let us assume that the distributions with negative means have zero variance and that the distributions with positive means are normal and have the variance $\sigma^2=4$.

Let the means and variances be:

$$\mu_1=\ldots=\mu_M=1; \quad \sigma_1^2=4$$

$$\mu_{M+1}=\ldots=\mu_{2M}=-1; \quad \sigma_2^2=0^1$$

[1] An example not quite so extreme is the following (1): The difference S–P between observed precipitation in the gauge-type S (new SMHI gauge) and gauge-type P

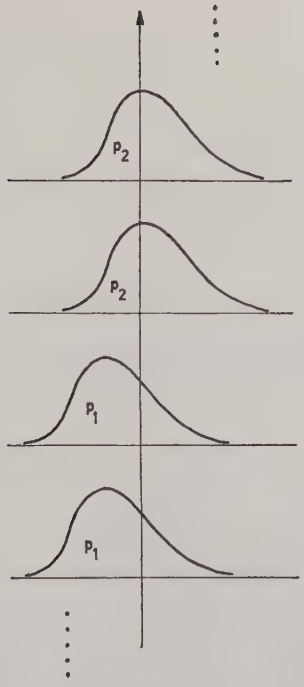


Fig. IV: 10.

The probability of a negative difference is again $>\frac{1}{2}$. In Wilcoxon's test what is the probability of rejecting the true null hypothesis in the negative direction with 2.5%'s level of significance, that is

when $b=6$ what is $P(T=0)=?$

when $b=7$ what is $P(T\leqslant 2)=?$

$$P(T_{\text{pos}}=0)=\sum_{\nu=0}^{b}\binom{b}{\nu}\left(\frac{1}{2}\right)^{b-\nu}\left(\frac{1}{2}\right)^{\nu}\left\{\Phi\left(-\frac{1}{2}\right)\right\}^{\nu}=\frac{\{1+\Phi(-\tfrac{1}{2})\}^{b}}{2^{b}}$$

$$b=6\Rightarrow P(T=0)=0.0784$$

(Pluvius) was studied. It showed that for small quiet rainfalls more precipitation is observed in P (perhaps because of an overly large collection area), thus S–P becomes negative. Because of the small amounts of precipitation the variances will be small too. On the other hand, when the rains are violent some drops will splash out of P because of its construction, and S–P then becomes positive. The variances may be larger as more precipitation is collected.

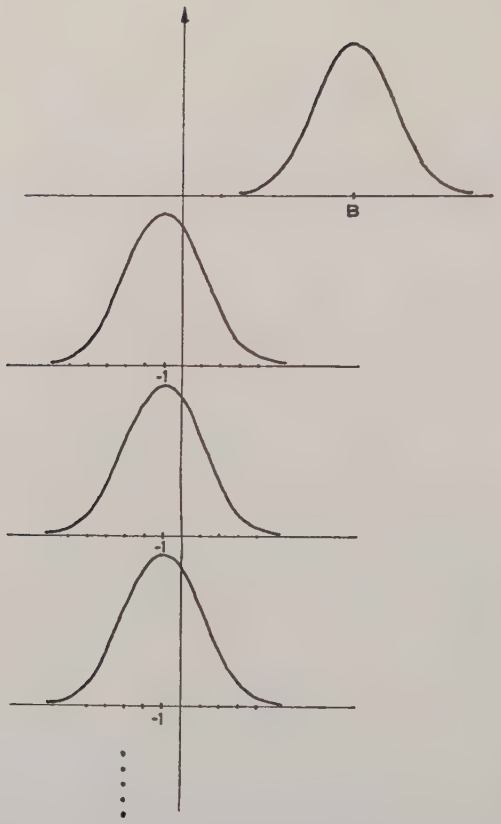


Fig. IV: 11.

$$P(T_{\text{pos}}=1)=\sum_{\nu=1}^{b}\binom{b}{\nu}\left(\frac{1}{2}\right)^{b-\nu}\left(\frac{1}{2}\right)^{\nu}\binom{\nu}{1}\left\{\frac{1}{2}-\Phi\left(-\frac{1}{2}\right)\right\}\left\{\Phi\left(-\frac{1}{2}\right)\right\}^{\nu-1}$$

$$=b\cdot\left(\frac{1}{2}\right)^{b}\left\{\frac{1}{2}-\Phi\left(-\frac{1}{2}\right)\right\}\sum_{\nu=1}^{b}\binom{b-1}{\nu-1}\left\{\Phi\left(-\frac{1}{2}\right)\right\}^{\nu-1}$$

$$=b\cdot\left(\frac{1}{2}\right)^{b}\left\{\frac{1}{2}-\Phi\left(-\frac{1}{2}\right)\right\}\left\{1+\Phi\left(-\frac{1}{2}\right)\right\}^{b-1}$$

$$P(T=2)=\sum_{\nu=2}^{b}\binom{b}{\nu}\left(\frac{1}{2}\right)^{b-\nu}\left(\frac{1}{2}\right)^{\nu}\binom{\nu}{2}\cdot 2\cdot P_{\nu}$$

where $$P_{\nu}=\int_{-\frac{1}{2}}^{0}\varphi(t)\{\Phi(-\tfrac{1}{2})-\Phi(-t-1)\}\{\Phi(-t-1)\}^{\nu-2}dt$$

$$\geqslant\{\Phi(-\tfrac{1}{2})\cdot\Phi^{\nu-2}(-\tfrac{3}{2})-\Phi^{\nu-1}(-1)\}\{\tfrac{1}{2}-\Phi(-\tfrac{1}{2})\}$$

Thus $P(T=2) \geq b(b-1)(\frac{1}{2})^b \{\frac{1}{2} - \Phi(-\frac{1}{2})\} [\Phi(-\frac{1}{2}) \{1+\Phi(-\frac{3}{2})\}^{b-2}$
$- \Phi(-1)\{1+\Phi(-1)\}^{b-2}]$

$b=7$ gives $P(T=0)=0.0513$

$P(T=1)=0.0526$

$P(T=2) \geq 0.0311$ and

$P(T \leq 2) \geq 13.5\%$

to be compared with the nominal 2.5%.

In the following two examples we will study the extreme case where only one distribution has a positive mean (and median), but where the average mean is still zero.

Example 11. Identical and normal distributions. (Fig. IV: 11).

$$\mu_1 = B, \quad \sigma_1^2 = 1$$

$$\mu_2 = \ldots = \mu_{B+1} = -1, \quad \sigma_2^2 = 1$$

$$P(T_{\text{pos}}=0) = \left(1 - \frac{b}{B+1}\right) \cdot \Phi^b(1)$$

$$\left.\begin{matrix} b=10 \\ B+1=1000 \end{matrix}\right\} \Rightarrow P(T_{\text{pos}}=0) = 0.99 \cdot (0.8413)^{10} = 0.1758$$

Example 12. Different distributions.
The distributions with negative means have zero variance.

$$\mu_1 = B, \quad \sigma_1^2 = 100(B-1)^2 \quad \text{(normal distribution)}$$

$$\mu_2 = \ldots = \mu_{B+1} = -1, \quad \sigma_2^2 = 0$$

When $b > 10$, H_0 will be rejected with probability 1. See table G in (18). When $b \leq 10$ it is easy to show that $P(T_{\text{pos}}=0) \geq 0.46$.

Discussion.—The null hypothesis examined by the Wilcoxon matched-pairs signed-ranks test apparently states that the total distribution of all differences (the sum of the distributions for all blocks in the population) is symmetric around the vertical axis through zero. In the sign test it only states that the average median is zero. When the distribution is symmetric Wilcoxon's test has greater power than the sign test, as it gives greater weight to "stronger" signs than to "weaker" signs.

In examples 5 and 6 the possible influences caused by deviations from unimodal symmetric distributions have been hinted at. Example 6 includes

the largest possible deviation relevant in this test. In practice the distribution will more often be in the neighbourhood of the distribution in example 5, where the distortion is slight.

The quantitative null hypothesis of zero average mean is tested when the average median and the average mean coincide, as, for example, when the distributions are symmetric and identical (except possibly for the mean). The examples illustrate that there may be very large deviations from the qualitative hypothesis although the quantitative hypothesis is true, especially when the distributions are very different.

Let us consider a pure qualitative situation where it is not possible to assign any numerical scores to the observations, but only to rank the two observations within a block, and then to rank the differences $1-2$ or $2-1$ according to their relative "strength".

Of course it is not sufficient for the Wilcoxon test to be applicable to have as many rank orders $1-2$ as $2-1$ in the population. There must be a sort of symmetry so that "strong" positive signs correspond to "strong" negative signs and so on.

For the sign test it was easy to define the qualitative null hypothesis, and qualitative additivity. They required average median $=0$ and constant difference between probabilities of $+$ and of $-$ from block to block respectively.

When Wilcoxon's test is to be applied the situation is unclear, as quantitative considerations enter the picture to some extent.

3.3.2. *More than two levels*

When there are more than two levels we use Friedman's two-way test based on ranks (8). For two levels this test is equivalent to the sign test, but the sign test is easier to handle. To test the hypothesis that the a factor levels having the observed rank sums: T_i, $i=1, 2 \ldots a$, do not differ, using b blocks, we compute:

$$\chi_T^2 = \frac{12}{ba(a+1)} \cdot \sum_{i=1}^{a} T_i^2 - 3b(a+1)$$

This statistic will be distributed approximately as χ^2, with $a-1$ degrees of freedom, as long as the number of rows and columns is not too small. Exact sampling distributions for small values[1] can be found in (8) and (18).

The null hypothesis tested says that within a block the a levels have identical distributions. Then the rankings within blocks are totally random

[1] For $a=3$, $b=2$ to 9, and $a=4$, $b=2$ to 4. When the values are larger the χ^2-distribution is thought to be a useable approximation.

and the rank in a cell can be considered as a random sample of size one from a discontinuous rectangular distribution having the possible values $1, 2 \ldots a$. The rank sums T_i will then be approximately normally distributed and have a known mean and a known variance (see below).

Friedman considers mostly quantitative variables, and when he has more than one observation per cell in some examples, he just takes the means and ranks these within every block. When the scale is ordinal his procedure would correspond to ranking all the observations in a block, taking the means of the ranks in every cell and finally, ranking these means. An alternative would be to extend Friedman's test to the case of more than one rank per cell. We will here make an extension, equivalent to that developed in chapter 2 for Cochran's Q-test, which is certainly more efficient than the method of ranking means.

Let the number of observations in the cells (subclasses) be r and denote the mean ranks by $\overline{T}_{ij} (i = 1, 2 \ldots a;\ j = 1, 2 \ldots b)$. The mean and variance in a discontinuous rectangular distribution having the values $1, 2 \ldots ar$, is: $(ar+1)/2$ and $(a^2r^2-1)/12$. We thus get the following relations:

$$E(\overline{T}_{ij}) = \frac{ar+1}{2}; \quad D^2(\overline{T}_{ij}) = \frac{a^2r^2-1}{12r} \cdot \frac{ar-r}{ar-1} = \frac{(a-1)(ar+1)}{12}$$

$$a \cdot D^2(\overline{T}_{ij}) + a(a-1) D^2(\overline{T}_{ij}) \varrho_{\overline{T}_{ij}\overline{T}_{i'j}} = 0; \quad \varrho_{\overline{T}_{ij}\overline{T}_{i'j}} = -\frac{1}{a-1}; \quad i \neq i'$$

The mean ranks for the factor levels are

$$\overline{T}_i = \frac{\sum_{j=1}^{b} \overline{T}_{ij}}{b}$$

and it is easy to see that

$$E(\overline{T}_i) = \frac{ar+1}{2}; \quad D^2(\overline{T}_i) = \frac{(a-1)(ar+1)}{12b}; \quad \varrho_{\overline{T}_i\overline{T}_{i'}} = -\frac{1}{a-1}; \quad i \neq i'$$

The central limit theorem then tells us that the $\overline{T}_i$'s are approximately normally distributed when a, b and r are not too small. From this it follows by applying Walsh's theorem referred to above (19), that:

$$\chi_T^2 = \frac{\sum_{i=1}^{a} \left(\overline{T}_i - \frac{ar+1}{2}\right)^2}{\frac{a}{a-1} \cdot \frac{(a-1)(ar+1)}{12b}}$$

is approximately distributed as χ^2, with $a-1$ degrees of freedom. A more convenient form for χ_T^2 is:

$$\chi_T^2 = \frac{12}{r^2ab(ar+1)} \cdot \sum_{i=1}^{a} T_i^2 - 3b(ar+1)$$

Putting $r=1$, we will arrive at Friedman's formula.

If the number of observations per treatment is r_j in block j, r is substituted by $\bar{r}=\sum_{j=1}^{b} r_j/b$ in the formula above.

Using the null hypothesis above, we need not presuppose anything except that the observations be randomly sampled. The hypothesis requires identical distributions within a block, but nothing is said about the form of these distributions and they can even differ from block to block. The strict assumptions included in the F-test are identical and normal distributions, the same in every block. Undoubtedly we have profited by using the rank test instead of the F-test when the measurements are quantitative. For qualitative measurements rank tests are usually considered as the only alternative. However we profit less when, like Friedman, we want to test the "block" factor too. Then the distributions have to be identical in both rows and columns.

Even if the above test offers a great advantage compared to the F-test, some difficulties remain. What happens, for example, to the test when the factor levels have equal means while the distributions differ in form? We have seen above, in 3.3.1, that this can cause trouble. What happens if there is interaction between blocks and treatments? Friedman (8) makes some comments on the possibility of different skewness and different variances but does not reach any general conclusion. More important is, I think, to consider the question of interaction. Concerning this Friedman says (8): "The method of ranks does not provide for testing 'interaction'. It is of the very nature of the method that it cannot do so. Without exact quantitative measurement, 'interaction', in the sense used in the ordinary analysis of variance, is meaningless." He also says: "... χ_T^2 may not be significant ... because the influence of the factor tested is different in direction for the different rows (blocks). In this connection, however, the general point should be emphasized that non-significant results do not establish the validity of the null hypothesis in the same way that significant results tend to contradict it."

Different influences in different blocks is what in this paper is meant by interaction. The null hypothesis used above requires that both main effect and interaction be zero. One would think that χ_T^2 would reject the hypothesis only when the main effect $\neq 0$, because when there is interaction

and zero main effect, the $\overline{T}_i$ tend to be more equal and thus give a small value to χ^2_T. However, as we are concerned with blocks sampled from a large population of blocks, the significance can be greatly overestimated (compare the Cochran Q-test in chap. 2). This is because the χ^2_T-test corresponds to an F-test with "mean square for within cells" as error term.

In order to be able to define zero main effect in the qualitative sense, let us consider the following ideal situation:

We are interested in the relations between a treatments for a population of blocks. In every block there is a certain rank order between the treatments. Taken over all the blocks in the population every treatment has its rank sum $T_i^{\text{Pop.}}$. The null hypothesis of zero main effect is then:

$$H_0\colon T_1^{\text{Pop.}} = T_2^{\text{Pop.}} = \ldots = T_a^{\text{Pop.}}$$

This does not leave out the possibility of interaction, which means that the ranks are not the same in every block.

If there is a risk of observing the wrong rank orders the situation is a little more complicated. Analogous to what has been said in connection with the sign test, the null hypothesis may not mean the same thing, as when there are no classification errors.

Fictitious numerical example.—Let us say that the population consists of 5 types of equally frequent blocks in a very large number. The true rank orders between the three treatments A_1, A_2 and A_3 for the block types are:

Block type no.	Treatment		
	A_1	A_2	A_3
1	1	2	3
2	1	2	3
3	3	1	2
4	2	3	1
5	3	2	1
$\Sigma_J T_J$	$k \cdot 10$	$k \cdot 10$	$k \cdot 10$

Below are given sampling distributions for χ^2_T when b, the number of blocks in the sample, equals 3. In the first column is given the distribution when we are sampling from the population without classification errors (i.e. we always observe the correct rank order). The distribution in the second column is for the case when there are equal probabilities of observing 1, 2, 3; 2, 1, 3 and 1, 3, 2 in the first block type, and the one in the third column is taken from Friedman (8), that is, when all possible rank orders are equally probable for every block.

χ^2_T	$P(\chi^2_T)$ without classification error	$P(\chi^2_T)$ with a certain type of classification error	$P(\chi^2_T)$ every possible rank order equally probable
0	0.096	0.0693	0.056
0.667	0.336	0.3760	0.416
2	0.312	0.2204	0.167
2.667	0.048	0.1298	0.167
4.667	0.120	0.1609	0.166
6	0.088	0.0436	0.028

If we use Friedman's table we will probably reject the hypothesis when $\chi^2_T = 6$. The probability of this is, in the first case, raised by 6 percentual units although the main effect is zero. A more extreme case is when we have only the block types

$$1, 2, 3 \quad \text{for} \quad J = 1, 2 \ldots B/2$$

$$3, 2, 1 \quad \text{for} \quad J = B/2+1, \ldots B$$

When $b=3$, this gives: $P(\chi^2_T = 6) = 2 \cdot (\frac{1}{2})^3 = 0.25$.

Let us say that there are M possible rank orders. When the probability is zero for ties, we have $M = (ar_J)!$; where, as before, r_J = the number of observations per treatment in block J. The probabilities of observing these rank orders in block J are:

$$p_{JK}; \begin{cases} J = 1, 2 \ldots B \\ K = 1, 2 \ldots M \end{cases}$$

with

$$\sum_K p_{JK} = 1$$

With the help of this we can compute expected values of mean ranks within the cells, $E(\bar{T}_{ij})$, and define additivity in the following way:

$$E(\bar{T}_{i1}) = E(\bar{T}_{i2}) = \ldots = E(\bar{T}_{iB}); \quad i = 1, 2 \ldots A$$

This definition, which includes the definitions used above, is the quantitative additivity applied to ranks. We have a special case of additivity when

$$p_{1K} = p_{2K} = \ldots = p_{BK}; \quad K = 1, 2 \ldots M,$$

which can be used when we want to construct examples with additivity without much effort.

The null hypothesis of zero main effect is formulated:

$$H_0: E(\bar{T}_1) = E(\bar{T}_2) = \ldots = E(\bar{T}_A)$$

However, it is easy to see that Friedman's test may tend to reject the hypothesis even when there is additivity, in the sense used above, and zero main effect. If in every block the rank orders (three treatments) are: 1, 2, 3, and 3, 2, 1, both with a probability of $\frac{1}{2}$, main effect and interaction are then zero. The probability is then 0.11 of obtaining at least 8 out of 10 blocks with one and the same rank order. If we obtain 8 blocks with the same rank order $\chi_T^2 = 7.2$, which is significant at the level 0.05. On the other hand when every block has only one possible rank order and all M rank orders are equally frequent in the population, then the sampling distribution of χ_T^2 will be the same as that of Friedman's even though there is interaction in our sense.

Friedman's test examines the hypothesis that in any observation every possible rank order is equally probable.

When the number, r, of observations per cell is large, we can, as has been discussed in chapter 2, compute a χ_T^2 for every block and then add these to a χ_T^2, with $b(a-1)$ d.f. In this way we test main effect, with $(a-1)$ d.f., as well as interaction, with $(a-1)(b-1)$ d.f., both with an "error term" corresponding to $\mathrm{MS}_{\text{within}}$. However, when r is large, it may be possible to use F-tests on ranks.

Below we shall write out a model for the observed ranks T_{ijl} of the lth replication of the ith treatment in block j. With the notations used above we get:

$$T_{ijl} = \frac{ar+1}{2} + \left\{E(\bar{T}_i) - \frac{ar+1}{2}\right\} + \{E(\bar{T}_{ij}) - E(\bar{T}_i)\} + \{T_{ijl} - E(\bar{T}_{ij})\}$$

Above $(ar+1)/2$ is the general mean if there is an equal number of observations in every cell, $E(\bar{T}_i) - (ar+1)/2$ reflects main effect and $E(\bar{T}_{ij}) - E(\bar{T}_i)$ interaction between blocks and treatments. The term $\{T_{ijl} - E(\bar{T}_{ij})\}$ denotes some random error and may be denoted by e_{ijl}. Possible differences between blocks will enter the model only in the interaction term, as the block means are constantly equal to $(ar+1)/2$. The validity of an F-test for main effects will depend on the variances of $E(\bar{T}_{ij}) - E(\bar{T}_i)$ in different treatments and the distributions of e_{ijl}. These distributions will be far from normal and will possibly contain heterogeneous variances; moreover, the e_{ijl} will not be independent. However, the interdependence between the e_{ijl}'s and the deviations from normality will decrease with increasing ar.

We have now constructed a qualitative model (that is, defined additivity and zero main effect) which is so general that Friedman's test is insufficient. However, using the F-test it is possible to test interaction and main effect separately. Therefore it would be of great value to find out when the F-test can be used on ranks.

In Friedman's case when all possible rank orders are equally probable it is clear intuitively that F-tests can be used if r is not too small, but it is in the case of interaction that F-tests would be of the greatest value. Simulation experiments are planned in Friedman's case as well as in some more general cases where different rank orders have different probabilities of occurring, and where interaction is present.

Friedman's test utilizes all possible permutations of ranks in the same way as the permutation tests discussed in part II utilize all possible permutations of the observed values in the blocks. In both cases the hypothesis tested includes zero main effect and zero interaction, but in the case of ranks the within-variances are known, and we can use χ^2 instead of F, which is used as an approximation of the permutation test.

Applications

Example 13. Comparative precipitation measurements.

In a comparison between 4 different types (P, SG, SN_1 and SN_2) of gauges for making precipitation measurements (1), 4 gauges of each type were used and emptied 2 times a day when it rained.

The following table, in which the P-gauges were numbered 2–5 and so on, gives a small portion of the primary data:

Date 1959	P				SG				SN_1				SN_2			
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
9/4 09	4.0	4.2	3.8	3.9	3.6	3.5	3.5	3.5	3.6	3.6	3.6	3.5	2.9	2.8	2.7	2.8
10/4 13	13.6	13.0	13.6	13.5	13.5	13.4	13.4	13.6	14.2	14.1	14.0	14.3	14.5	14.8	14.7	14.4
⋮																
2/5 09	1.6	1.5	1.4	1.5	1.2	1.2	1.4	1.1	1.6	1.5	1.6	1.5	0.5	0.6	0.4	0.4
⋮																

It was obvious from the data that the gauges collected different amounts of precipitation and the main interest was therefore to estimate the magnitudes of these differences and to find out how they varied with the amount of precipitation. In this context, however, we shall study the problem of testing the hypothesis that the types of gauges give the same results.

These above observations are really quantitative, but, because of their very great number (the investigation was carried on for two years), it was considered desirable to use a more rapid method of analysis than the analysis of variance. It is also apparent that the variances within type of gauge increase with the amount of precipitation, thus the distributions vary from block to block, which is generally considered to invalidate the F-test. Therefore Friedman's test was considered feasible, and it was carried through by ranking the sums for gauges of the same type. It would have been possible to rank the individual measurements and use the above mentioned extension to Friedman's test, but it would have caused more work and also a considerable increase in the frequency of ties. Besides, a highly significant result was already obtained with the sums.

Let us now see what hypothesis we have tested by using Friedman's test. The gauges of each type are considered representative, that is as randomly sampled from the (hypothetical) population of gauges. The hypothesis with Friedman's test is that in every block every possible rank order is equally probable. (When the rains are randomly sampled from a population of rains, this hypothesis is not as restrictive as it may sound—see discussion above.) It means that not only every type of gauge but also every individual gauge in the 4 populations of gauges will give the same result on the average for all possible blocks.[1] The observed differences have to be completely random. Thus a significant result may be obtained, because:

a) the types of gauges differ on the average;
b) the types of gauges give different results for different types of rain (interaction);
c) the variations between gauges of the same type differ from type to type.[2]

Earlier it has been discussed that interaction may influence the risks of both kinds of error when the model is mixed. Although measurements have been made for two years, it is possible that, for example, violent rain-storms have been fewer than in normal years. An analysis of variance test would take this into account by "testing against interaction". It

[1] With a new sample of gauges for every rain it would have been sufficient to assume identical distributions for the 4 populations of gauges.

[2] If the problem can be formulated something like: "Do these gauges show the same result, that is does it matter which I use?", the hypothesis to be tested is that of Friedman's (average effects and interaction are zero). If, on the other hand, there is the problem of comparing annual precipitation one may wish to test the hypothesis of zero average effects.

would test whether or not the differences between types of gauges are large compared to the differences between gauges of the same type. When the data are quantitative the advantage of using Friedman's test instead of the F-test is often neutralized by the loss in flexibility. However, an F-test is in this example complicated by the fact that we have to use a linear combination of 3 sums of squares in the denominator.

The hypothesis that no differences between gauges of the same type occur may be tested more exactly. It was tested using Friedman's test separately for each type of gauge. If there is no interaction the test is valid even though both "treatments" and blocks are random. Possible differences will probably be caused by differing collecting areas, which means that the quantitative differences increase with the amount of precipitation but that the rank order is unchanged. Thus it is a reasonable assumption that there exists a quantitative but not a qualitative interaction.

As a numerical example the test has been computed on monthly sums for one year. See gauges no. 2–17 in (1), appendix p. 392. The following is obtained:

$$\left.\begin{array}{ll} \text{Individual gauges ranked:} & \chi_T^2 = 69.11 \\ \text{Types of gauges ranked:} & \chi_T^2 = 29.58 \end{array}\right\} \chi^{2(3)}_{.0005} = 17.7$$

An analysis of variance on the measurements has also been carried out, and the results recorded in the following table:

Sources of variation	SS	D.f.	MS	E(MS)
Between types of gauges (M)	516.38	3	172.13	$\sigma_e^2 + \sigma_{RE}^2 + 4\sigma_{RM}^2 + 6\sigma_E^2 + 8\sum_{i=1}^{4} \alpha_i^2$
Between gauges within type (E)	66.43	12	5.54	$\sigma_e^2 + \sigma_{RE}^2 + 6\sigma_E^2$
Between months (R)	106 804.73	5	21 360.94	$\sigma_e^2 + \sigma_{RE}^2 + 16\sigma_R^2$
Interaction ($R \times M$)	105.86	15	7.06	$\sigma_e^2 + \sigma_{RE}^2 + 4\sigma_{RM}^2$
Interaction ($R \times E$)	87.06	60	1.45	$\sigma_e^2 + \sigma_{RE}^2$

The hypothesis that the types of gauges are equivalent on the average, H_0: $\Sigma_i \alpha_i^2 = 0$, is tested with

$$F = \frac{\text{MS}_M}{\text{MS}_E + \text{MS}_{RM} - \text{MS}_{RE}} = 15.44$$

According to Satterthwaite's well known approximation (15), one obtains 3 and 21 d.f. for the test variable. As the variances apparently vary from block to block (increasing with the precipitation) it is doubtful whether the mean squares are uncorrelated, but the significance is too high to be

affected, viz.: $F^{3;20}_{.0005}=9.20$. The hypotheses H_0: $\sigma^2_E=0$ and H_0: $\sigma^2_{RM}=0$ are both rejected at the level 0.0005. Thus there is (ignoring the small risk of having committed an error of type I) interaction between monthly sums and types of gauges, average differences between types and within types. As χ^2_T reacts on all these causes there is no mystery why it gave such a large numerical value.

As a comparison, an analysis of variance on ranks has been made with the following result:

Sources of variation	D.f.	MS	F
Type of gauge (M)	3	522.17	15.81
Gauges within type (E)	12	31.19	33.18
Interaction ($R \times M$)	15	2.77	2.95
Interaction ($R \times E$)	60	0.94	

The F-values have been computed in the same way as they have been computed from the quantitative values and the results are about the same. The F-value for gauges within type has greatly increased, a consequence of their small but persistent differences.

Example 14. Comparing psycho-physical reactions

In the psycho-physical application described in part III, paragraph 4.3, it is doubtful whether the scale of measurement is more accurate than an ordinal scale. Below the analysis of variance table is reproduced and the result of Friedman's test given.

Analysis of variances table.

Sources of variation	D.f.	MS	F	$F_{.05}$	$F_{.01}$	$F_{.005}$
Subjects	9	0.89				
Tones	3	1.48	5.29	2.96	4.60	5.36
Interaction	27	0.28	1.65	1.77		
Residual	40	0.17				

Result of Friedman's test

Individual observations ranked: $\chi^2_T=12.55$; $\chi^{2(3)}_{.01}=11.3$

Cell means ranked: $\chi^2_T=5.97$; $\chi^{2(3)}_{.10}=6.25$

Thus the F-test and Friedman's test for individual ranking gave significance at the level 1%. Apparently it is possible to incur considerable

power losses by ranking cell means instead of individual observations. The good correspondence between the two first mentioned tests is due to the fact that the interaction is unimportant. Individual ranking gave rise to many ties, but this only makes the test more conservative (8).

3.4. More than one factor

If we add a factor, in which we are not especially interested, but which can be considered as a classification factor, no new complication is introduced. We simply obtain more blocks and, possibly, an increased interaction between treatments and blocks.

If the experiment contains several treatment factors of special interest difficulties arise. Friedman (8) has used his test method on both factors in a two-way design, but as mentioned above this requires further restrictive assumptions. As a rule one wants to study the interaction between the two factors, and this can often not be managed by the nonparametric tests.

Sometimes, however, nonparametric tests for the interaction can be used, see, for example, Mood (13) and Wilson (22). Wilson's method is based on the median test and utilizes a partitioning of the total χ^2 into (with two factors) one χ^2 for rows, one for columns and one for interaction. Rao (14) has shown that such a partitioning of χ^2 and of the degrees of freedom is possible. First of all, the median for all the observations (scores) is determined. Thus the test is applicable when all observations are on the same ordinal scale. If, of all n observations, n_a are above and n_b below the median we expect $(n_a/n)\cdot r_{ij}$ observations above the median in a cell having r_{ij} observations when the rows are equivalent, the columns are equivalent and there is no interaction. Denoting the number of observations above the median in cell ij by ${}_af_{ij}$, and the number below the median by ${}_bf_{ij}$, the total χ^2 can be written:

$$\chi^2_{To} = \sum_{i=1}^{a}\sum_{j=1}^{b}\left\{\frac{({}_af_{ij} - (n_a/n)\cdot r_{ij})^2}{(n_a/n)\cdot r_{ij}} + \frac{({}_bf_{ij} - (n_b/n)\cdot r_{ij})^2}{(n_b/n)\cdot r_{ij}}\right\}$$

When the hypothesis of the same average median for every row is true, we expect $(n_a/n)\cdot r_{i.}$ observations above the median in every row $(r_{i.} = \sum_{j=1}^{b} r_{ij})$. Thus we obtain:

$$\chi^2_A = \sum_{i=1}^{a}\left\{\frac{({}_af_{i.} - (n_a/n)\cdot r_{i.})^2}{(n_a/n)\cdot r_{i.}} + \frac{({}_bf_{i.} - (n_b/n)\cdot r_{i.})^2}{(n_b/n)\cdot r_{i.}}\right\}$$

In the same way we obtain for columns:

$$\chi^2_B = \sum_{j=1}^{b}\left\{\frac{({}_af_{.j} - (n_a/n)\cdot r_{.j})^2}{(n_a/n)\cdot r_{.j}} + \frac{({}_bf_{.j} - (n_b/n)\cdot r_{.j})^2}{(n_b/n)\cdot r_{.j}}\right\}$$

TABLE IV: 11. *Number of values below* ($_bf$) *and above* ($_af$) *median for each combination of subjects and tones.*

Subject no.		Tone A	B	C	D	Sums
1	$_bf$	0	0	2	0	2
	$_af$	2	2	0	2	6
2	$_bf$	1	0	0	0	1
	$_af$	1	2	2	2	7
3	$_bf$	2	2	1	1	6
	$_af$	0	0	1	1	2
4	$_bf$	1	0	1	0	2
	$_af$	1	2	1	2	6
5	$_bf$	2	2	1	2	7
	$_af$	0	0	1	0	1
6	$_bf$	1	0	0	0	1
	$_af$	1	2	2	2	7
7	$_bf$	2	2	1	1	6
	$_af$	0	0	1	1	2
8	$_bf$	2	0	2	0	4
	$_af$	0	2	0	2	4
9	$_bf$	1	0	1	0	2
	$_af$	1	2	1	2	6
10	$_bf$	2	1	2	0	5
	$_af$	0	1	0	2	3
Sums		14	7	11	4	36
		6	13	9	16	44

The chi-square value for interaction, χ_I^2, is then computed as the difference:

$$\chi_I^2 = \chi_{To}^2 - \chi_A^2 - \chi_B^2$$

The numbers of degrees of freedom for the sums of squares are respectively $(a-1)(b-1)$; $ab-1$; $a-1$; and $b-1$, and it is obvious how the tests should be made.

As individual cells (not, for example, column sums) have been used the r_{ij}-values must not be too small. In this case it may be possible to make F-tests on ranks which will utilize more information than can be had by solely recording "above" or "below" the median.

To illustrate Wilson's method let us apply it to the psycho-physical experiment reviewed in part III and used as example 14 in chapter 3.3. If the median is chosen to be something between 20.0 and 20.3, we will get $n_a = 44$ and $n_b = 36$. The number of values above and below the median are shown in table IV: 11.

The expected values $\frac{36}{40} = 0.9$ and $\frac{44}{40} = 1.1$, for the total χ^2 are of course too small in order to give a good approximation to the χ^2-distribution, but the example can be used to illustrate the computations, and to make some comparisons with the other tests that have been carried through. For rows and columns the expected values are higher; thus a poor approximation will influence chiefly the test for interaction.

We obtain the following table:

Sources of variation	D.f.	χ^2	$\chi^2_{.80}$	$\chi^2_{.01}$	$\chi^2_{.005}$
Tones	3	11.717		11.3	12.8
Subjects	9	23.434		21.7	23.6
Interaction	27	20.607	20.7		
Total	39				

The resulting probabilities, except the one for interaction, are about the same as when the F-tests are applied to scores. However, these χ^2-tests are always performed like an F-test in the fixed model. If we construct an F-test for tones with interaction as error term we obtain:

$$F = \frac{\frac{1}{3} \cdot 11.717}{\frac{1}{27} \cdot 20.607} = 5.12$$

to be compared with $F = 5.29$, when the analysis was made on the scores.

Example 15. Comparing shapes of drinking-glasses.

In an investigation of the functions of every-day glasses a study was made into what shapes of glasses made the washing and drying most convenient. Five types of glass-shape were compared. The glasses were used for drinking milk and juice and washed in two makes of dish-washing machines. Every glass-type was placed in each of 5 places in the machines. In the first machine it was possible to wash 5 glasses of every type at the same time, in the second only 3. There is then one factor of interest, "types of glass", two crossed classification factors, "machines" and "drinks", and one classification factor, "places", nested within machines. After the experiment every glass was attributed one of the scores 1, 2, 3, 4, where 4 meant perfectly clean. Of course these scores cannot be treated as quantitative measurements. One possibility would be to rank the glasses according to the scores in all the 20 blocks, but as there are 15 to 25 glasses, but only 4 possible scores in a block, there would be a lot of ties. A better method is to compare perfectly clean glasses (formerly with the score 4) with not perfectly clean glasses (formerly bearing scores of 1, 2 or 3), and not to distinguish between degrees of dirtiness. The scale has then become nominal.

REFERENCES

1. Andersson, T.: Further studies on the accuracy of rain measurements. Arkiv för Geofysik, band 4, nr. 16 (1963), 359–393.
2. Brownlee, K. A.: Statistical theory and methodology in science and engineering. Wiley, New York (1960).
3. Cochran, W. G.: The comparison of percentages in matched samples. Biometrika 37 (1950), 256–266.
4. — Some methods for strengthening the common χ^2-tests. Biometrics 10 (1954), 417–451.
5. Cox, D. R.: Planning of experiments. Wiley, New York (1958).
6. Dixon, W. J. and A. M. Mood: The statistical sign test. Jasa 41 (1946), 557–566.
7. Fisher, R. A. and F. Yates: Statistical tables for biological, agricultural and medical research. Oliver and Boyd, London (1949).
8. Friedman, M.: The use of ranks to avoid the assumption of normality. Jasa 32 (1937), 675–701.
9. Gabrielsson, A.: Betoningsupplevelse i korta tonsekvenser. Planering och förundersökningar. Unpublished lic. thesis in psychology, Uppsala (1964).
10. Kruskal, W. H. and W. A. Wallis: Use of ranks in one-criterion variance analysis. Jasa 47 (1952), 583–621.
11. Kruskal, W. H.: A nonparametric test for the several sample problem. AMS 23 (1952), 525–540.
12. Mann, H. B. and D. R. Whitney: On a test of whether one of two random variables is stochastically larger than the other. AMS 18 (1947), 50–60.
13. Mood, A. M.: Introduction to the theory of statistics. McGraw-Hill, New York (1952).
14. Rao, C. R.: Advanced statistical methods in biometric research. Wiley, New York (1952).
15. Satterthwaite, F. E.: An approximate distribution of estimates of variance components. Biometrics 2 (1946), 110–114.
16. Scheffé, H.: The analysis of variance. Wiley, New York (1959).
17. Seeger, P.: Olika modeller för variansanalys av ett autentiskt exempel. Unpublished lic. thesis, Uppsala (1962).
18. Siegel, S.: Nonparametric statistics for the behavioural sciences. McGraw-Hill, New York (1956).
19. Walsh, J. E.: Concerning the effect of intraclass correlation on certain significance tests. AMS 18 (1947), 88–96.
20. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics Bull. 1 (1945), 80–83.
21. — Probability tables for individual comparisons by ranking methods. Biometrics 3 (1947).
22. Wilson, K. V.: A distribution-free test of analysis of variance hypotheses. Psychological Bulletin 53 (1956), 96–101.